1-1-2013

Association of Protein Helices and Assembly of Foldamers: Stories in Membrane and Aqueous Environments

Shaoqing Zhang
University of Pennsylvania, zhangsh@sas.upenn.edu

Association of Protein Helices and Assembly of Foldamers: Stories in Membrane and Aqueous Environments

Abstract
Solvents play an important role in association and assembly of molecules. Here we studied solvent effects on proteins and organic chemicals in different contexts. First, X-ray crystal structures show that helix dimers in membrane- and water-soluble proteins have distinct behaviors in packing and sequence selection. Transmembrane dimers are stabilized by compact packing and hydrogen bonding between small residues. Meanwhile, water-soluble dimers utilize hydrophobic residues for packing irrespective of the size of the interface and tight dimers are rare. Secondly, we apply the results learned above to a complex system in which a designed protein binds to single-walled carbon-nanotube in aqueous environments. Previous designs of the hexameric helical bundles utilized leucine and alanine residues to make two distinct helix-helix interfaces. Our molecular dynamics simulations showed that the alanine-comprising interface is much more labile than the leucine-comprising one. This result can be interpreted by the scarcity of tight soluble helix dimers as mentioned above. Thus more stable modular helix-helix interfaces have to be employed to design peptides binding to carbon-nanotubes with higher affinities. Lastly, we describe a serendipitous discovery of the crystalline framework structure by an amphiphilic triarylamide foldamer. Foldamers are peptide-like polymers of non-natural monomers arranged in defined sequence and chain length that are able to adopt protein-like secondary and tertiary structures. In contrast with traditional metal-organic and organic frameworks, which exploit strong directional coordination and hydrogen bonding for assembly in organic solvents, the crystal herein is built up from a combination of noncovalent hydrophobic, hydrogen-bonded, and electrostatic interactions in aqueous solution. The structure is in honeycomb geometry with each cubicle as a truncated octahedron. A new supramolecular synthon, in which hydrogen bonding and π-π stacking are encompassed, was discovered in the crystal structure. Through NMR experiments we probed the oligomeric states of the foldamer in the early stages prior to crystallization. The hierarchic crystal structure was discussed in terms of supramolecular synthons in crystal engineering.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Physics & Astronomy

First Advisor
William F. DeGrado

Subject Categories
Atomic, Molecular and Optical Physics | Bioinformatics | Chemistry

ASSOCIATION OF PROTEIN HELICES AND ASSEMBLY OF FOLDAMERS: STORIES IN MEMBRANE AND AQUEOUS ENVIRONMENTS

Shaoqing Zhang

A DISSERTATION

in

PHYSICS AND ASTRONOMY

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

_____

William F. DeGrado, Professor, Pharmaceutical Chemistry

Interim Graduate Group Chairperson

_____

Randall D. Kamien, Professor, Physics

Dissertation Committee

William F. DeGrado, Professor, Pharmaceutical Chemistry
Mark Goulian, Professor of Biology & Physics
Prashant K. Purohit, Associate Professor of Mechanical Engineering and Applied Mechanics
Ravi Radhakrishnan, Associate Professor of Bioengineering
Wei Tong, Associate Professor of Pediatrics

ASSOCIATION OF PROTEIN HELICES AND ASSEMBLY OF ORGANICS: STORIES IN MEMBRANE AND AQUEOUS ENVIRONMENTS

To my parents

ACKNOWLEDGEMENTS

Don't cry because it's over, smile because it happened.

— Dr. Seuss

Upon finishing my Ph.D. years, I must express my gratitude to many people. They have helped me a lot in my studies and life, and have cared about me and my growth in every aspect.

It is my fortune to be one of Bill's students and to work with him. He is a scientist with enormous ingenuity and incredible amiability. He always has great and novel projects in a diverse array of fields for me. The ideas are always challenging but great fun to work on. He has let me set the pace and has never pushed me for immediate results on any project. I appreciate his patience and trust very much. Whenever discussing my projects with him, his deep insight and broad perspective have always inspired me. His great passion in science has been infectious to me. Working with Bill, I have felt self-motivated and joyful to work on science. Besides this, he is also a great cook. I like very much the food he made. If we had dinners at his house and could not finish the food, I felt very lucky to take some home.

I want to thank Mark, Prashant, Ravi, and Wei for serving on my committee and granting me their support. It was a great experience for me to take courses from Mark, Prashant and Ravi. Their teaching showed the fun of science in different fields. After years I can still remember the scenes and feel the happiness in class. Wei showed me the elegance of the signaling mechanisms in hematology, which greatly sparked my curiosity

in research in this field.

I am very grateful to the people in the DeGrado family. All the people in the lab, the current and former members, have given great help and an enjoyable life to me. Brett, my baymate and buddy, has helped me a lot in my work and life. We have shared a lot of feelings together. The third Texan, Gabe, has given me so many favors ~~to~~ in getting my computer running and has taught me a lot about everything from his perspectives. We are proud to be Texans. Michelle has helped me greatly with the Ph.D. defense talk along with Brett and Gabe. The notes you put down for my talk were tremendously helpful. Nate opened the door of crystallography for me. We two have always been hotel mates through the years, and he and his wife Susan have helped my family a lot. Yibing, my office neighbor, has always talked about everything with me and is also my close collaborator in science. Jo, our lab manager, has helped so much with my peptide synthesis work, and has always spurred me to be a greater scientist. I have received a lot of help with my lab life from Kathleen, who was the first lab member I met before I joined the lab and when I was TAing her. Jenny has helped me a lot with the setup for computational work and thanks a lot to her for the collaboration on our metalloprotein project. Bruk is the molecular biology master in the lab and has helped me such a lot with molecular cloning. Gözde, my partner of various projects, has shown me a fresh and fascinating world of metalloorganic chemistry, the knowledge of which has helped greatly with my organic framework research. Jun taught me how to optimize protein expression and his diligence in research always motivates me. Marco, my collaborator far from Italy, has shared with me a lot of great time since he came to UCSF one month ago. Gaby has teamed up with me for the crystallization job and has helped me a lot with

protein expression. I always enjoyed talking with Zac. Lastly, I want to thank a special member of our lab, Susan, Bill's the other half, who has helped greatly with lab parties and retreats and other lab affairs by offering places and cooking for us.

I also want to thank the former members who have overlapped me in the lab. They have provided me very important help for my scientific startup in the DeGrado lab. Gevorg, taught me a lot on protein design and related computer coding throughout the time since I joined the lab as a rotation student. He and his wife Keila has helped my family a lot. My wife and I have been excited with the birth of their baby girl, with a beautiful Armenian name Nairi. Jason has helped me so much with the buildup of rudimentary knowledge in protein modeling and we share the hobby of book collection. He and his wife Melissa, who has a Chinese origin, have had a great time with my family on Chinese learning and Chinese food. Cinque has taught me a lot about science and the life outside, and has introduced one of his great Chinese friends to me. Ivan, the former lab manager, spoiled me at Penn by his constant conversations with me when I was a layman in everything in lab. Jade has helped me a lot on HPLC at UCSF. When we were moving from Penn, we helped each other a lot. Yao helped me a lot within and without the lab. Dan has also help me a lot with learning coding when I was rotating in the lab and he was busy with the preparation for his Ph.D. defense. I thank Chaim for collaboration on the helix dimer projects, and he is also a master in computational biology. Paul, Rudy, Yongho, Ilan and Michael always shared great stories and jokes on everything with me. I have enjoyed great times with John, Graham, Shalom, Geronda and other lab members. It was also nice meeting former members with whom I had no overlap: Greg, Alessandro, Vik, Karin, Byan, Jonna, Scott and other people on various

occasions.

I want to thank my other academic mentors along the way in my career. Firstly, I am very grateful to Dr, Su, who was the advisor for my master's study. He took me into the field of biophysics, and cared about my life in every aspect. With his support, he was able to enter Penn for a doctor's degree. I want to thank Margaret, who taught me how to do MD simulations and helped me understand the meanings of life in many ways. I also would like to thank Wenying, who showed me the way biological experiments could be done. Dr. Weglein was my first mentor in the United States. I want to thank my undergraduate advisor Dr. Chao, who has offered me great encouragement when I felt weak. And thanks to Dr. Guo for his trust in my capabilities in scientific research.

I would like to thank my friends in Houston, in Philadelphia and in San Francisco. With them I can feel there are always people I can get help from when I get desperate. I want to thank the classmates and colleagues I worked with at University of Houston. They have shaped me into my current personality to some extent. The folks from Peking University in Houston have given me utmost great help, and I have received so much pleasure by staying in contact with them. Two of them, Jin and Qianghua, have become my life-long friends. Lei, my former college classmate, has helped me a lot in Houston. Ming taught me a lot and gave me a lot of help when we were working with Dr. Su. Two special friends I got when I was studying in different labs or groups are Fang and Enxiu. Fang taught me to use *vi*, the editing tool I have been using in Linux, and he was trying to protect me whenever I got frustrated during my very first year in United States. Enxiu gave me enormous care in lab and in life when I started doing experiments for the first time in a biology lab at Penn. Di also helped me a lot in science and in life at Penn. In

San Francisco, Wei, who is always staying with me, has given me a lot of help. Huayong, another former college classmate and friend, now at Berkeley, always cares for me in every way. There are a lot of people who have helped greatly in various occasions, and I am very grateful to them also.

Finally I want to thank the most important people of my life, my family. My parents tried their best efforts and sacrifice to raise me and give me the best education. This thesis is dedicated to them. I want to thank my wife, who has provided the most precious care and forbearance during my most difficult times. My sister is ever-lastingly supportive of me with every decision I have made.

ABSTRACT

ASSOCIATION OF PROTEIN HELICES AND ASSEMBLY OF FOLDAMERS: STORIES IN MEMBRANE AND AQUEOUS ENVIRONMENTS

Shaoqing Zhang

William F. DeGrado

Solvents play an important role in association and assembly of molecules. Here we studied solvent effects on proteins and organic chemicals in different contexts. First, X-ray crystal structures show that helix dimers in membrane- and water-soluble proteins have distinct behaviors in packing and sequence selection. Transmembrane dimers are stabilized by compact packing and hydrogen bonding between small residues. Meanwhile, water-soluble dimers utilize hydrophobic residues for packing irrespective of the size of the interface and tight dimers are rare. Secondly, we apply the results learned above to a complex system in which a designed protein binds to single-walled carbon-nanotube in aqueous environments. Previous designs of the hexameric helical bundles utilized leucine and alanine residues to make two distinct helix-helix interfaces. Our molecular dynamics simulations showed that the alanine-comprising interface is much more labile than the leucine-comprising one. This result can be interpreted by the scarcity of tight soluble helix dimers as mentioned above. Thus more stable modular helix-helix interfaces have to be employed to design peptides binding to carbon-nanotubes with

higher affinities. Lastly, we describe a serendipitous discovery of the crystalline framework structure by an amphiphilic triarylamide foldamer. Foldamers are peptide-like polymers of non-natural monomers arranged in defined sequence and chain length that are able to adopt protein-like secondary and tertiary structures. In contrast with traditional metal-organic and organic frameworks, which exploit strong directional coordination and hydrogen bonding for assembly in organic solvents, the crystal herein is built up from a combination of noncovalent hydrophobic, hydrogen-bonded, and electrostatic interactions in aqueous solution. The structure is in honeycomb geometry with each cubicle as a truncated octahedron. A new supramolecular synthon, in which hydrogen bonding and π-π stacking are encompassed, was discovered in the crystal structure. Through NMR experiments we probed the oligomeric states of the foldamer in the early stages prior to crystallization. The hierarchic crystal structure was discussed in terms of supramolecular synthons in crystal engineering.

Table of Contents

List of Tables

List of Figures

Chapter 1

1.1 Introduction

The structure of molecules at the atomic level is vital to developing a fundamental understanding of various types of physical and chemical interactions. X-ray crystallography, NMR and molecular dynamics simulations are the three complementary means to determine molecular structures at the atomic level. The former one determines molecular structures in the crystalline state; the other two elucidate solution structures. The structures in these two states are closely correlated. The crystalline state is usually obtained by evaporation of the solvent to concentrate the solute. The crystal structure is thus governed by the configuration of the solute in its solvated state.

There are two main categories of proteins according to their solvation properties: membrane- and water-soluble proteins. A major class of membrane proteins is transmembrane proteins, hydrophobic regions of which are inserted in a lipid bilayer of around 30 Å in thickness. Lipid molecules consist of a polar head-group and a nonpolar aliphatic chain. The hydrophobic core, which is formed by self-association of the nonpolar chains, provides a fluid-like solvating environment for transmembrane proteins. This is very stringent solvation conditions compared to the three-dimensionally isotropic aqueous one. Physical and chemical interactions that drive protein folding in membrane and aqueous environments are distinctly different from each other. Water-soluble proteins usually require hydrophobic interactions to fold into well-shaped native states. Transmembrane proteins, which can be solubilized in the lipid bilayer, require compact

packing and polar interactions between their secondary structures. Within the membrane environment, association is driven largely by tight and efficient packing as well as hydrogen bond formation [1, 2]. Compact packing can be attained by small residues, because large residues usually have higher side-chain entropies. Due to the low dielectric constant inside the hydrophobic core of the lipid bilayer, polar interactions can be very strong compared to the aqueous environment. However, large polar residues have high solvation energy in the lipid bilayer. Thus the folding of transmembrane proteins is subject to a trade-off between solvation energy and van der Waals packing.

To explore the folding of transmembrane and soluble proteins, we can study their X-ray crystal structures. Structure in the crystalline state is the epitome of an ultra-high concentrated state. The folded structures of transmembrane and soluble proteins have been found to contain recursively occurring structural and sequence motifs. Usually a sequence motif determines a structural motif. The motifs for association between basic secondary structures – helix-helix interaction, have been investigated extensively. In transmembrane proteins, several motifs have been identified, including GxxxG [3], SxxSSxxT and SxxxSSxxT [4], and QxxS [5]. The small polar residue asparagine induces helix-helix association [6, 7]. In soluble proteins, the leucine zipper [8] and Ala-Coil [9] motifs are well known. However, these studies on the sequence motifs were specific to certain proteins. To discover globally utilized motifs in transmembrane and soluble proteins, the complete database of protein X-ray structures were examined. We clustered helix dimer structures and then analyzed the sequence profile, i.e., we first identified the structure motifs and then the sequence motifs. We found that helix dimers

in transmembrane proteins tend to pack more tightly than in soluble proteins. The sequence motifs and their geometric configuration for tight structure motifs are determined. Small polar residues are popular in the transmembrane tight motifs. Meanwhile, we observed that there are no strong biases for loose transmembrane helix dimer motifs, other than that the interface should be packed with hydrophobic amino acid sidechains. By contrast, motifs in water soluble proteins show a statistical bias towards the use of hydrophobic residues to pack between helices, irrespective of the type of motif. When the motif is looser, larger hydrophobic residues are used. Therefore studies on crystal structures can well illustrate the separate manners of protein folding in membrane and aqueous milieus. This is the content discussed in Chapter 2.

Due to versatile functions of proteins, they can play a key role in biotechnological applications, where proteins are adsorbed to surfaces, especially at liquid-solid interfaces. They can be used in enzymatic activity, tissue engineering, and bioelectrochemical reactions. Protein binding to single-walled carbon nanotubes (SWCNTs) and graphene is of great interest, because these two allotropes of carbon possess intriguing electronic properties. In Chapter 3 we describe the binding between a *de novo* designed protein and SWCNT. Our design was inspired by the domain swapped dimer (DSD) protein [10], in which a large linear void was created by its oligomerization (Figure 1A). The proteins were designed by keeping the helix-helix interfaces in the DSD protein and building new sequences for hydrophobic interior and hydrophilic exterior to bind to SWCNTs in aqueous environments (Figure 1B) [11]. One design, called HexCoil-Ala, was shown experimentally to bind SWCNTs [11]. Crystallization of the mixture of HexCoil-Ala and

SWCNTs generated only crystals for tetrameric protein helical bundles (Figure 1C) [11]. We measured binding affinity of HexCoil-Ala to SWCNTs and determined the configurations of HexCoil-Ala with SWCNTs by molecular dynamics (MD) simulations. MD simulations are a powerful tool to elucidate atomic-level structures of systems that are too large for NMR studies and too labile for crystallization. Our MD simulations showed that the structure of HexCoil-Ala on SWCNTs is hexameric but deviates significantly from the design. The designs contain two types of interface: leucine zipper and Ala-Coil. The leucine zipper motif as a helix-helix interface displays great robustness in structure, while the Ala-Coil motif is very labile. HexCoil-Ala has a hydrophobic core to wrap SWCNT and an aqueously solvated surface. Due to non-directionality of hydrophobic interaction and circular symmetry of SWCNTs, HexCoil-Ala proteins undergo a constant rotational motion around SWCNTs. The helix-helix association of HexCoil-Ala belongs in aqueous environments as the volume of SWCNTs is much smaller than that of aqueous solvent. As we have found in Chapter 2, helix dimers in soluble proteins do not tend to put small residues glycine and alanine at their interfaces because of their low hydrophobicity. Thus the Ala-Coil interface is energetically less favorable than a motif rich in large hydrophobic residues in aqueous conditions. At high concentrations, HexCoil-Ala will tetramerize to lower free energy and thereafter form crystals. MD simulations helped us elucidate limitations of the previous designs and facilitate better future designs for SWCNT-wrapping proteins.

More generally, solvents have a strong influence on the modes of association and assembly of organic molecules. Chemists have employed the physico-chemical principles

of solvent effects to achieve various reactions. One of the burgeoning fields is metal-organic frameworks (MOFs), the crystalline compounds in which rigid organic molecules are coordinated by metal ions. MOFs have been applied in gas storage, catalysis and sensing [12]. MOF crystals are usually obtained by mixing organic molecules and metal ions in nonpolar solvents and subsequently evaporating the solvents. Organic molecules with more than one polar functional group can make coordination complexes with multiple metal ions that are not hydrated in nonpolar environments. When the solvents are evaporated, the coordination complexes are connected as a large network. Because the coordination geometry of metal ions is fixed at a specific oxidation state, the network is a regular infinite framework existing as a MOF crystal. In Chapter 4, we describe a serendipitous discovery by crystalizing one infinite framework assembled by hydrophobic interactions between organic molecules in aqueous environments. The organic molecules are amphiphilic: the hydrophilic regions interact with water and salt ions; the hydrophobic regions self-associate to become shielded from water. The crystal displays a honeycomb geometry with each cubicle as a truncated octahedron. Hydrophobic interactions between trifluoromethyl (-CF_3) groups are located at the vertices of the framework; π-π interactions between 1,3-diaminobenzene groups are placed at the edges. Structural rigidity of the organic molecule is conferred by binding of cadmium(II) ions. Thus the properties of solvents determine the assembly modes of organic molecular frameworks. We are interested in understanding the assembly pathway from solution state to crystalline state, the investigation of which is rarely attempted in the MOF field. NMR experiments were conducted to examine the oligomerization states of the organic molecules at high and low concentrations in aqueous conditions. At

concentrations near those used for crystallization the foldamer is in an equilibrium between monomers, dimers and higher order oligomers. Therefore the framework formation is a process of concentration-dependent transitions. Investigation of structures and pathways of the assembly can help us understand the feasibility of designing materials in the form of frameworks in aqueous environments.

The studies above tell us that solvent effects on association and assembly of organic molecules and macromolecules are essential.

1.2 Figures

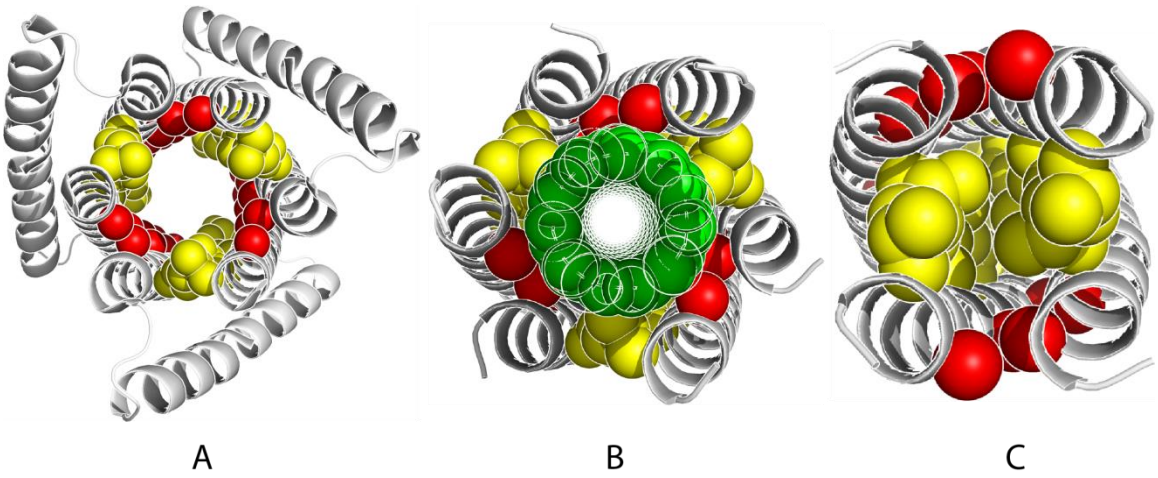


Figure 1. The structure of the DSD protein (A), HexCoil-Ala model with SWCNT (B) and HexCoil-Ala tetrameric crystal structure (C). The leucine zipper and Ala-Coil interface are color yellow and red, respectively. SWCNT is colored green.

1.3 References

1. Bowie, J.U., *Solving the membrane protein folding problem.* Nature, 2005. **438**(7068): p. 581-9.
2. White, S.H. and W.C. Wimley, *Membrane protein folding and stability: physical principles.* Annu Rev Biophys Biomol Struct, 1999. **28**: p. 319-65.
3. Russ, W.P. and D.M. Engelman, *The GxxxG motif: a framework for transmembrane helix-helix association.* J Mol Biol, 2000. **296**(3): p. 911-9.
4. Dawson, J.P., J.S. Weinger, and D.M. Engelman, *Motifs of serine and threonine can drive association of transmembrane helices.* J Mol Biol, 2002. **316**(3): p. 799-805.
5. Sal-Man, N., D. Gerber, and Y. Shai, *The identification of a minimal dimerization motif QXXS that enables homo- and hetero-association of transmembrane helices in vivo.* J Biol Chem, 2005. **280**(29): p. 27449-57.
6. Choma, C., et al., *Asparagine-mediated self-association of a model transmembrane helix.* Nat Struct Biol, 2000. **7**(2): p. 161-6.
7. Zhou, F.X., et al., *Interhelical hydrogen bonding drives strong interactions in membrane proteins.* Nat Struct Biol, 2000. **7**(2): p. 154-60.
8. O'Neil, K.T. and W.F. DeGrado, *A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids.* Science, 1990. **250**(4981): p. 646-51.
9. Gernert, K.M., et al., *The Alacoil: a very tight, antiparallel coiled-coil of helices.* Protein Sci, 1995. **4**(11): p. 2252-60.
10. Ogihara, N.L., et al., *Design of three-dimensional domain-swapped dimers and fibrous oligomers.* Proc Natl Acad Sci U S A, 2001. **98**(4): p. 1404-9.
11. Grigoryan, G., et al., *Computational design of virus-like protein assemblies on carbon nanotube surfaces.* Science, 2011. **332**(6033): p. 1071-6.
12. Farrusseng, D., *Metal-organic frameworks : applications from catalysis to gas storage*. 2011, Weinheim: Wiley-VCH. xxii, 392 p.

Chapter 2

A New Dictionary of Helix-Helix Interactions in Membrane and Soluble Proteins

2.1 Overview

Alpha helices are a basic unit of protein secondary structure, and the interaction between two helices is therefore crucial to understanding tertiary and higher-order folds. Structural and sequence motifs can help by precisely describing a specific type of helix-helix interaction and highlighting the crucial residues. Moreover, comparing subtle variations in these motifs between membrane and soluble proteins can shed light on the different constraint faced in each environment and elucidate the complex puzzle of membrane protein folding. Here, we demonstrate that soluble helix pairs cluster into a small number of distinct geometries, as has previously been shown for transmembrane helix pairs. Similarly placed amino acids in a given helix pair show different interactions for helix-helix association in membrane and aqueous milieu. We also analyze the sequence profiles of each cluster to find statistically significant amino acid biases and establish their important contributions to dimer stability. We further characterize known and novel packing geometries that feature distinct interhelical topologies. Investigation of these clusters will greatly improve our understanding of the sequence-structure relationship in transmembrane and soluble helical proteins. They also provide a structural basis for molecular modeling and rational templates for protein design.

2.2 Introduction

In water-soluble proteins approximately 35% of all protein residues are in the α-helical conformation [1], making it by far the most common regular secondary structure element.

Moreover, membrane proteins are almost exclusively α-helical bundles, with the excpetion of the β-barrels found in the outer membrane of Gram negative bacteria and mitochondria. Over 30% of the homologous superfamilies described in CATH are comprised mainly or entirely of alpha helices [2]. These domains are found in both soluble (SOL) and transmembrane (TM) proteins, and carry out a wide range of biological functions.

Since the first transmembrane protein was crystallized in 1984 [3], the folding mechanism of these proteins has gradually become clearer [4], but much work remains. They are estimated to make up 20-30% of open reading frames in known genomes [5], and are overwhelmingly alpha helical, containing one or multiple membrane-spanning helices. Specific interactions between helices play a critical role in the function, assembly and oligomerization of these proteins [6]. However, TM proteins represent only 2% of deposited structure [7] due to experimental challenges in crystallization. Computational and bioinformatics-based study of helix-helix interactions will therefore assist us in understanding the folding behavior of helical TM proteins.

An open question is whether helices from TM and SOL proteins are the same in the way they interact with each other and contribute to overall protein structure. This can be broken down into individual properties, such as helical content, length, as well as dimer properties like interhelical distance and crossing angle. It is already known that a small subset of SOL helix-helix pairs are structurally homologous to TM pairs and have similar properties, even though the overall distributions for SOL dimers are quite different from those of TM dimers [8]. Here, we investigate the full range of SOL helix-helix interactions and compare them to those found in TM proteins, focusing on the interplay

between sequence and structure. To do this, we extend the approach used previously for characterizing TM dimers [9] to a larger database of TM dimers with more strict criteria and compare the results with dimers from water soluble proteins.

Analysis of sequences derived from helix-helix dimers propels our understanding of helix-helix interactions. The most extensively studied TM helix dimer is the model system, Glycophorin A (GpA) [10, 11]. Each helix of GpA contains two Gly separated by three amino acids, known as the GxxxG motif [12], which play a key role in dimerization. The GxxxG motif is highly overrepresented in the sequences of TM proteins [13], and has been well-characterized structurally. GxxxG-containing dimers tend to have a parallel, right-handed geometry, compact helix-helix packing [11] and stabilizing interhelical backbone hydrogen bonds. Comprehensive characterization via a variety of biophysical and biochemical methods has established the GxxxG motif as an important framework of TM helix-helix interaction [14]. Gly can be commonly replaced by another small residue, such as Ala or Ser in this motif [13, 14]. The Ala-Coil [15] or GxxxxxG motif is another prevalent sequence motif found in membrane protein families [16]. Other sequence motifs have also been identified, which depend on hydrogen bonds or weak polar interactions, and include derivatives of the small-residue motifs mentioned above [4, 17-29].

However, a systematic study of sequence-structure relationships on the scale of the whole protein structure database using structural bioinformatics is still lacking. Here we extract helix-helix pairs from high-resolution, non-homologous TM and SOL proteins from the protein data bank (PDB), and cluster them based on geometric similarity. This is one of the first such comprehensive analyses of the clusters of soluble helix dimers. We contrast

the relative frequencies of each cluster in both environments and identify specific conformations that are unique to one or the other. Notably, sequence profiles can differ between the TM and SOL datasets, even for geometrically identical clusters. We also analyze the interactions of statistically enriched residues in seven clusters of TM helix dimers and in their structural counterparts in SOL dimers. Characterization of these sequence and structural motifs will contribute greatly to our understanding of the folding of helical proteins and aid both in structure prediction and de novo design.

2.3 Results

2.3.1 Helix Pairs Assume a Limited Number of Geometries

The range of the interhelical distance the helix dimers own is larger than in the previous study by Walters and DeGrado (WD) [9]. We expanded the definition of helix dimers to include pairs with interhelical distance up to 14 Å. Meanwhile, more stringent criteria for inclusion in clusters: previously we used a RMSD cutoff of 1.5 Å, and a minimum length of 10 residues on each helix, which is changed to a more stringent 1.25 Å and 12 residues, respectively. The inclusion of relatively long inter-helical distances and stringent clustering criteria give different clustering results comparing with WD analysis.

We find 20 clusters of TM and 15 clusters of SOL helix pairs whose population is shown in Figure 1, which include 51.1% and 50.7% of the total 1725 TM and 5085 soluble dimers, respectively. The WD study of helix-helix interactions in transmembrane proteins found about a quarter of the number helix pairs [9]. These include clusters in all four canonical geometries (parallel and anti-parallel, right- and left-handed), with a wide range of interhelical distances. This demonstrates that the grouping of dimers into

discrete clusters is a general feature of helix-helix interactions, and is not limited to a particular environmental subset.

We have 48.9% of TM dimers and 49.3% of SOL dimers that are not clustered. We examine why the dimers failed to get clustered. Unclustered pairs generally have a larger interhelical distance than the clustered ones (Supplementary Figure S1). More importantly, we have identified the largest cluster TM Cluster 1, which was ascribed to TM Cluster 6 in WD study due to a small interhelical distance and a loose RMSD cutoff. We are also able to capture an intermediate and several small TM clusters with an interhelical distance larger than 11.5 Å: Clusters 7 (AL), 10 (PL), 12 (AL), 15 (PL) and 17 (AL). Therefore a large interhelical distance cutoff in our current analysis gives rise to 2 new top clusters which own more than 5% of population.

The greater resolution and larger number of TM protein structures in the current study allowed us to more clearly define clusters than in previous studies. We have a new database containing 893 helix dimers extracted from 58 unrelated proteins. The top 5 clusters found in the WD analysis all appear within the top 7 TM clusters presented here (Table 1). In both the WD and current studies, each of these clusters has a population larger than 5% of the TM dimer library, allowing for statistically meaningful sequence analysis. The top 5 clusters in the WD analysis and the top 7 in the current study occupy 81.8% and 66.6% of the clustered population, respectively.

Helix dimers from a set of non-homologous SOL structures were also clustered. A total of 2761 dimers were extracted from 765 proteins. The pairs fit into 15 geometrically unique clusters. Similar to the TM clusters, the 15 clusters are large enough for sequence

analysis, and the top 7 SOL clusters contain more than 5% of the population, totaling 73.8% of the clustered dimers (Figure 1). They are structurally similar to the top 7 TM clusters (Table 2). In this article, we will focus on the top 7 SOL clusters and how they compare to their transmembrane equivalents.

We compare TM and SOL clusters by analyzing representative helix-helix dimers (centroids) from each cluster and discovering sub-segments of these helix-helix dimers that are the most structurally similar. This is accomplished by structurally aligning each 12-residue window (24-residues if you count both helices) of one TM cluster centroid to each 12-residue window of a SOL cluster centroid (see Methods). Surprisingly, this results in direct matches between the seven most populated clusters, that is the top 7 TM clusters and their 7 SOL counterparts. Two special cases exist where a cluster from one dataset is equally close to two clusters from the other dataset. SOL Cluster 5 is close to TM Clusters 1 and 6 and TM Cluster 7 is close to SOL Clusters 6 and 10 (Table 2). From this analysis, TM and SOL helix dimers tend to share remarkably similar geometry.

2.3.2 Geometric Trends

The dimers break down into parallel and anti-parallel, left-handed and right-handed groups. The geometries can be distinguished by the value of helix crossing angle Ω: ($-90^\circ < \Omega < 0$) for right-handed parallel (RP), ($90^\circ < \Omega < 180^\circ$) for right-handed antiparallel (RA), ($0 < \Omega < 90^\circ$) for left-handed parallel (LP), ($-180^\circ < \Omega < -90^\circ$) for left-handed antiparallel (LA). Together with interhelical distance, these two geometric parameters allow us to distinguish the different clusters (Figure 2). While the range of parameters is similar for both TM and SOL dimers, the distributions are weighted somewhat differently.

This reflects the fact that the relative sizes of the clusters are unique in each library (Figure 2). For instance, the largest TM and SOL clusters are geometrically the same at this resolution, but account for only 21.7% of clustered TM dimers, compared to 27.3% of clustered SOL dimers.

As expected, the top TM and SOL clusters reside in the most populated regions of parameter space, and matching clusters are close together (Figure 2). The LA region is the densest, containing three TM and two SOL clusters, including the largest ones. One cluster from each library is in both the RP and LP regions, while there are two of each in the RA region. More detailed distinctions within each region and the packing geometries of specific clusters will be discussed below.

2.3.3 Residue Preference

One important component of our analysis is the characterization of specific residues with statistically significant frequency at certain positions within the helices. The residues are postulated to make an increased contribution to dimer stability by compact packing or H-bonding. Within the membrane environment, association is driven largely by tight and efficient packing as well as hydrogen bond formation [30]. On the other hand, the folded state of soluble proteins is largely dictated by hydrophobic effects, and the residue preference of helix bundles and coiled coils has been studied in detail [31].

TM and SOL proteins have different background frequencies for each of the 20 amino acids (Supplementary Table S1), and this is accounted for when calculating which residues are over-represented (See Methods). The occurrence of Phe, Ile, Leu, Met, and Trp in the TM database is at least 1.5 times as high as in SOL due to the hydrophobic

nature of TM helices. The occupancy of Asp, Glu, Asn, Gln, Lys and Arg in SOL is at least 2.5 times as high as in TM due to their high energy cost of insertion in membrane milieu.

Using the structural alignments, we are able to align the sequences within each cluster and find positions at which specific residues are statistically significant. The propensity is defined as the ratio between the observed and expected (or background) frequencies. Significant residues are defined to have a propensity larger than 1.5 and a P-value less than 0.05.

In the SOL database, Leu is highly over-represented at biased positions [31]. Asp, Glu, Lys, Gln and Arg also appear at biased positions. Unlike its prominent role in membrane proteins, there are no Gly residues at biased positions in SOL dataset. Ser and Thr are frequently found at biased positions in TM dimers. Interestingly, Asn also appears at biased positions in TM clusters because it is small, without charge, can readily form hydrogen bonds and can be important in the folding of helical TM proteins [32, 33]. Small residues (Gly, Ala, Ser, Cys, Thr and Asn) allow for excellent packing, and the frequency with which each is used shows a nice correlation between residue size and interhelical distance.

Certain positions have silent mutations to physiochemically similar amino acids, such as those observed in small positions in the GxxxG motif [34]. Therefore, in addition to single amino acid biases, we examined the average propensity of similar amino acids. It helps greatly identify important packing residues in TM and SOL clusters.

2.3.4 Left-handed Antiparallel Clusters

TM Clusters 1 and 6 and SOL Cluster 1: TM Cluster 1 is the biggest TM cluster and comprises of 21.7% of the clustered pairs. It was not found in the WD analysis. Its structural counterpart is SOL Cluster 1, which is the largest SOL cluster with a population of 27.3%. Their centroid-to-centroid RMSD is 0.59 Å, and their crossing angles and interhelical distances are very close (Table 2 and Figure 2). SOL Cluster 1 is a prototypical coiled coil according to the characterized geometric parameters [31]. In its heptad repeats, positions *a* and *d* are occupied mainly by hydrophobic residues Val, Leu, Ile and Met.

TM Cluster 6 is has a crossing angle is very close to that of TM Cluster 1, but the inter-helical distance is much smaller (Table 2). It corresponds to Cluster 1 in WD classification. It is the known Ala-Coil motif with small residues occupying positions *a* in the heptad repeats (Figure 3). It holds 8.4% of the family. TM Cluster 6 has a larger RMSD with SOL Cluster 1 than TM Cluster 1 (Table 2).

TM Cluster 7 and SOL Clusters 6 and 10: TM Cluster 7 is the third left-handed antiparallel cluster, the other newly discovered TM cluster. Its crossing angle is close to those of TM Clusters 1 and 6. The inter-helical distance is the largest among the three. It possesses 6.2% in population. It has two structural matches in SOL clusters: Clusters 6 and 10. Their populations are 7.9% and 5.5%, respectively. They both have a large RMSD with TM Cluster 7. SOL Cluster 10 has a slightly larger interhelical distance but a much smaller crossing angle than Cluster 7 (Table 2). They have Leu, Ile, Met, and aromatic residues on positions *a* and *d* in their heptad repeats.

2.3.5 Left-handed Parallel Clusters

TM Cluster 4 and SOL Cluster 2: TM Cluster 4 is the only left-handed parallel cluster in the top 7. It has a population of 7.8%, and was assigned as Cluster 4 in WD classification. It owns a small crossing angle of 12.2° but a large inter-helical distance (Table 2). It has no sequence preference for packing. Its SOL structural counterpart, SOL Cluster 2, holds a larger population of 10.0% and is the 2nd biggest SOL cluster. SOL Cluster 2 has a larger crossing angle but a smaller inter-helical distance (Table 2). As in the wide clusters SOL Clusters 6 and 10, Leu, Ile, Met, and aromatic residues act as packing residues.

2.3.6 Right-handed Antiparallel Clusters

TM Cluster 2 and SOL Cluster 3: TM Cluster 2 has 7.1% percent of the clustered population. It corresponds to Cluster 2 in the WD classification. With the same handedness but the opposite orientation with GpA, it has a crossing angle close to that of the latter (-34.6°) and a close inter-helical distance (Table 2). There is one GxxxG motif on one helix and one generalized GxxxG motif on the helix where the second small residue is Thr. The two GxxxG motifs do not have direct packing interaction. Interestingly, the Thr residue in the GxxxG motif and the other three small polar residues (Asn, Ser or Thr) interact on the opposite flank of the GxxxG motif (Figure 3).

The structural match of TM Cluster 2 is SOL Cluster 3, which has a population of 9.5%. SOL Cluster 3 has a larger inter-helical distance and a wider crossing angle (Table 2). These two clusters have a large RMSD (1.83 Å). Val, Leu, Ile, Met and aromatic residues pack at the interface. In SOL Cluster 3 has a wider crossing angle, which can accommodate larger packing residues.

TM Cluster 5 and SOL Cluster 4: TM Cluster 5 has a large inter-helical distance and a narrow crossing angle. Its population is 7.9%. In WD classification, it was appointed Cluster 5. SOL Cluster 4 is its structural counterpart, which has a crossing angle very close to that of TM Cluster 5, and a slightly smaller inter-helical distance (Table 2). They have a very small RMSD of 0.53 Å. SOL Cluster 4 has 7.1% of population. As in SOL Cluster 3, Val, Leu, Ile, Met and aromatic residues are packing residues.

2.3.7 Right-handed Parallel Clusters

TM Cluster 3 and SOL Cluster 5: TM Cluster 3 was ranked as Cluster 3 in the WD classification and corresponds to the extensively-studied GxxxG motif. It has a population of 7.6%. Its crossing angle is very close to that of GpA (-34.6°). While the inter-helical distance is small, it is larger than the 6.60 Å found in GpA (Table 2). As found before, TM GxxxG motifs are asymmetrically packed. As shown in Figure 3, TM Cluster 3 has one GxxxG motif on one helix in the same way as GpA and on the other helix the GxxxG motif occurs toward the middle of the interface. An additional small residue on the second helix packs in between the small residues in the GxxxG motif on the first helix. Two larger residues with this small one and the GxxxG motif on the same helix generate a ridge for the small residues on the first helix to pack in. It is a "knobs-into-holes" packing configuration.

The structural counterpart of TM Cluster 3 is SOL Cluster 5 with a population 6.4%. It has a slightly wider crossing angle and a slimly larger interhelical distance (Table 2). The RMSD is as small as 0.65 Å. It is the GxxxG motif in soluble proteins [35]. Different from the traditional view, there is a GxxxG motif only on one helix. The small residues

can also substituted by Val. Shown in Figure 3, it is also a "knobs-into-holes" configuration: two packing patches comprised of large hydrophobic residues form a ridge, which is tightly docked by the small residues in the GxxxG motif on the first helix.

2.4 Discussion

TM and SOL clusters employ similar residues for helix dimerization. As shown in Supplementary Figure S2, hydrophobicity files of SOL clusters show regular patterns of polar and apolar residues on the positions along the sequences. They correspond to spatial water-exposed hydrophilic residues and the buried hydrophobic residues at the interface. There is little fluctuation in the hydrophobicity files of TM clusters on the positions along the sequence. The only exceptions are on the positions with small residues for packing, where Ala, Ser and Thr are the interfacial residues (in TM Clusters 2, 3 and 6). In TM Clusters the residues both facing the membrane milieu and the interface are hydrophobic, so the hydrophobicity profiles are flat. When TM and SOL clusters have close interhelical distances, they employ similar hydrophobic interactions for packing.

In SOL Cluster 3, the small residues in the GxxxG motif can be substituted with Val by sequence analysis. However, it does not take place in the all three small-residue comprising motifs, TM Clusters 2, 3and 6. Val can contribute more hydrophobic force for helix-helix dimerization than small residues. Because small residues Gly and Ala have low hydrophobicity, they are rarely found for helix-helix association in SOL clusters. Meanwhile, small residues confer compact packing for TM helix dimers. Small polar residues Ser, Thr and Asn appear in several close TM clusters. They also form hydrogen-bonding, which is very important for the association of TM helices. When the interhelical

distance is larger (> 9.0 Å), in TM clusters there is no sequence with high propensities. Hydrophobic residues are packing at the helix-helix interfaces due to their predominant presence of in TM helices. However, they do not have tight packing as in the small-residue comprising motifs.

TM and SOL clusters have different relationships between interhelical distance and crossing angle. Left-handed antiparallel clusters can demonstrate this well (Supplementary Figure S3). TM Clusters 1, 6 and 7 have close crossing angles, and interhelical distances span a large range from 8.33 Å to 11.55 Å (Table 2). Meanwhile, the three SOL Clusters in this category have a broader distribution of crossing angles. Crossing angle is thus one important factor for soluble helix dimerization. Soluble proteins rely heavily on hydrophobic interaction to fold themselves. SOL Cluster 6 and 10 are both structural counterparts of TM Cluster 7. SOL Cluster 6 has a larger population than Cluster 10, because its wider crossing angle can better facilitate packing of large hydrophobic residues. The structural matches between TM Cluster 4 and SOL Cluster 2 (LP), between TM Cluster 2 and SOL Cluster 3 (RA), between TM Cluster 3 and SOL Cluster 5 (RP) all shows a wider crossing angle in the SOL clusters (Table 2 and Figure 2).

Helix-helix association is also affected by other factors, e.g., length of the TM patch [36]. Investigation of the clusters will help greatly our understanding of the folding and structure of helical proteins, quantifying broad structural trends which will be useful in structure prediction and design.

2.5 Experimental Procedures

2.5.1 Dataset selection

The Orientation of Proteins in Membranes (OPM) database [37] was used as the source for helical TM proteins. We obtained a list of all structures available as of July 4, 2013. To ensure accurate analysis, structures with X-ray resolution lower than 3.2Å were removed from consideration. From the remaining structures, we used the PISCES server [38] to cull at the PDB ID level for a maximum sequence homology of 30%. This resulted in a list of 97 representative structures, from which helix-helix pairs were derived. For the soluble database, a query was executed on the PDB as of February 9, 2012 for all structures classified in CATH [2] as "mainly alpha" and containing only protein. These were matched against the PDB-TM database [39] and any TM proteins were removed. This list was also culled using the PISCES server to a maximum of 30% sequence identity. In order to keep the size of the dataset computationally tractable, only structures with a maximum resolution of 2.0Å were kept, resulting in 765 proteins. For all soluble structures, the biological unit was downloaded from the PDB.

We extracted the helical regions from the selected structures using the definitions of the TM segments in the OPM or the HELIX records in the PDB header information for soluble proteins. In order to ensure that these definitions were correct, the annotated regions were filtered to exclude helical breaks or sharp kinks (defined with a loose cutoff: $-130^\circ < \varphi < -20^\circ$ and $-90^\circ < \psi < 30^\circ$). They were also extended by up to 4 residues on both the N- and C-terminal sides if the positions meet a stricter definition of helicity ($-90^\circ < \varphi < -35^\circ$; $-70^\circ < \psi < 0^\circ$). This helped to join soluble helices that otherwise might have been counted separately.

2.5.2 Creating the pair library

Two heuristic criteria were used to determine whether a given pair of helices was interacting. First, the minimum distance between the helical axes was required to be no more than 14 Å; second, the mean inverse distance was required to be at least 0.065 $Å^{-1}$ over a 12-residue window (see "Window Selection and Alignment" below for a definition of this quantity). Both of these were intended to be generous, as low specificity would merely result in a larger fraction of dimers which cannot be clustered, while low sensitivity would negatively impact our ability to detect and characterize real trends.

Although the overall structural libraries were filtered to reduce sequence homology, individual proteins often contain multiple copies of one or more subunits, resulting in several identical helix pairs. In order to remove this additional source of redundancy polypeptide chains with identical sequences were assigned to a "chain group," which allowed us to identify and remove duplicate dimers. Two helices can either come from the same chain (a Type I pair), different chains, both belonging to the same chain group (Type II), or separate chains that also belong to disparate chain groups (Type III). The final helix pair library contains 1725 TM dimers (1402 Type I, 153 Type II and 170 Type III) and 5085 soluble dimers (4343 Type I, 657 Type II and 85 Type III).

2.5.3 Window Selection and Alignment

To be able to align pairs, we used a distance map representation of each dimer. Briefly, the inverse distance between each Cα atom on one helix and every Cα atom on the other is stored in a matrix. (Residues more than 25 Å apart are given a value of 0.) We selected a twelve-residue segment from each helix, chosen so that we captured the maximum

amount of interaction for a given pair. Interaction strength was determined by averaging the interfacial distance map over a 12-residue window on each helix (the "mean inverse distance"), as calculated using Equation 1:

$$M = \frac{1}{n^2} \sum_{i=a}^{a+n-1} \sum_{j=b}^{b+n-1} x_{ij} \quad (1)$$

where M is the interaction strength, n is the window size (here 12 residues), a and b are the starting residues of the window on each helix, respectively, and xij is the value of the distance map for residues i and j, i.e. the inverse of the distance between the Cα atoms of residues i and j (in Angstroms) or zero if they are more than 25 Å apart. M was maximized by varying a and b over all possible values, from 1 to L-n+1, where L is the length of the particular helix. Since residues that are closer together in three dimensions have a larger entry in the distance map, this picks out the twelve residues on one helix that are closest to twelve residues on the other. Moreover, because of the inverse weighting, this emphasizes each residue's nearest neighbors, with the distances between the end of one helix and the far end of the other being less important.

We used MaDCaT [40] to conduct all-vs.-all searches of the two dimer libraries. Interactions are not always symmetrical along the length of a helix, with six residues on either side of the point of closest approach –some are 'V'-shaped rather than 'X'-shaped. Thus had we merely compared the twelve-residue windows to each other directly, we would have missed pairs that otherwise have the same geometry. We therefore searched each query window against the library of whole pairs, as extracted above. We limited the searches to a maximum of 10,000 hits each, which in practice exhausted all possible alignments within our clustering threshold.

2.5.4 Structural Clustering

Examining the alignments calculated by MaDCaT, we chose a 1.25 Å RMSD cut-off for clustering as an appropriate balance between sensitivity and specificity. We used the same 12-residue windows described above; windows which overlapped by six residues or more on either helix were considered identical and clustered together, while windows with smaller overlaps are treated separately. (This allows the total number of alignments to be greater than the number of unique pairs.) To cluster the pairs, we computed all possible sub-threshold alignments to each window. The window with the largest number of alignments from unique, previously unclustered pairs was selected as the next centroid. All matching windows were assigned to that cluster and removed from consideration for further rounds. This process was then repeated until none of the remaining windows matched at least ~1% of the associated database (18 pairs for TM and 55 pairs for soluble).

We found 20 clusters of TM helix pairs, including 999 alignments from 882 unique pairs (51.1% of the database). From the SOL database, 15 clusters were extracted containing 2757 soluble alignments from 2576 unique pairs (50.7%). HELANAL [41] was used to determine helical axes for the calculation of geometrical properties, including crossing angle and interhelical distance of the aligned windows in each cluster.

2.5.5 Comparing Clusters

For each centroid, we determined the 15-residue window that is most populated by members of that cluster. To compare clusters, we then used MaDCaT to find the best possible alignment of 12 residues between each pair of centroids approximate to those

regions. This information allowed us to identify the most closely related clusters from different sets.

2.5.6 Sequence Analysis

We used the structural alignments generated by MaDCaT for each cluster to create sequence alignments. Briefly, each centroid pair was renumbered so that the C-terminal residue of the centroid window would be residue 100. Each member of a cluster was then renumbered to match the centroid numbering, such that residues with the same number correspond in the structural alignment. The numbers of observations for every amino acid type were computed for each position in each cluster. These were compared with the expected amino acid probabilities observed in the each library overall. We use the overall observed frequency of amino acids in our TM dimer database as the expected frequency. TM and SOL background frequencies are listed in Supplementary Table S1. If this ratio was greater than or equal to 1.5 and the number of counts had a p-value ≤ 0.05 by the binomial test, the residue was considered significant or biased. Because rare amino acid with a very small count can satisfy the two criteria, the total counts of observation on each position are also considered in the sequence-structure analysis. Hydrophobicity profiles were calculated based on the normalized consensus scale [42].

2.6 Acknowledgements

Chaim A. Schramm, Daniel W. Kulp and Alessandro Senes are the co-authors of this chapter. I thank Brett Hannigan and Gabriel Gonzales for technical help and Ilan Samish for useful discussions. This work was supported by NIH grant R37GM54616.

2.7 Figures

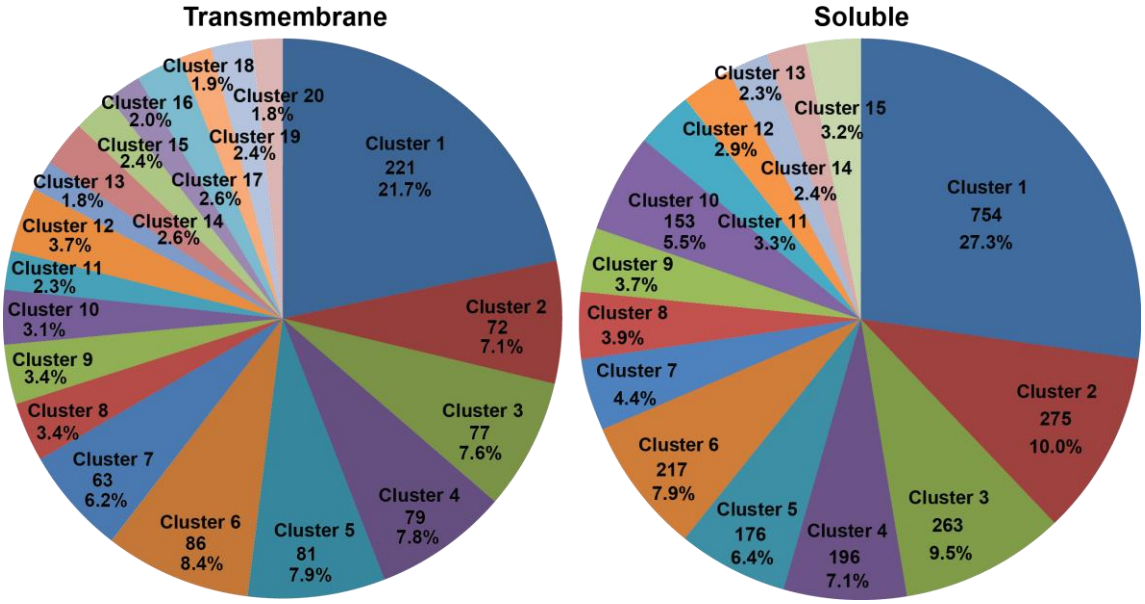


Figure 1. Pie chart showing factions of the 20 TM and 15 SOL clusters. Number of pairs is shown for the top 7 clusters in TM and SOL database.

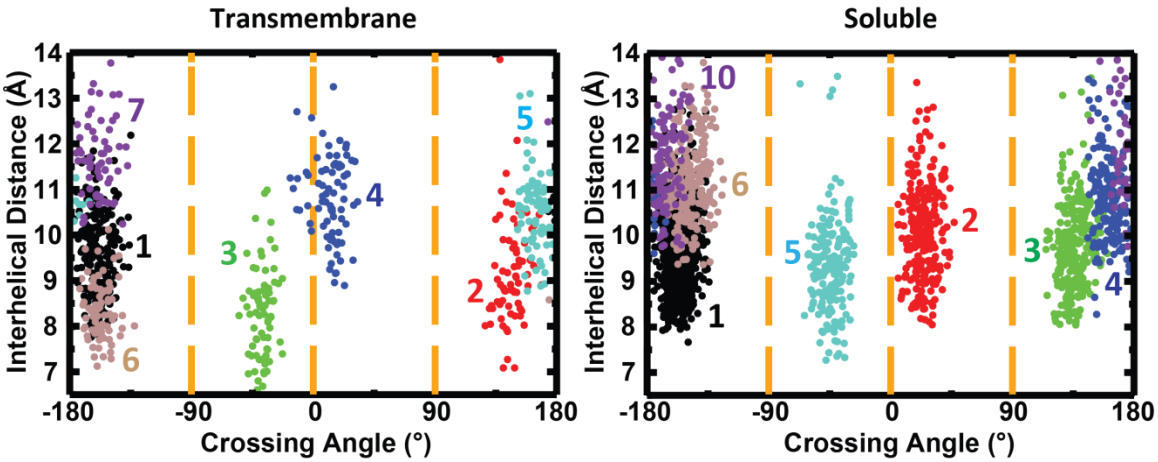


Figure 2. Distribution of crossing angle and interhelical distance of top 7 TM and SOL clusters.

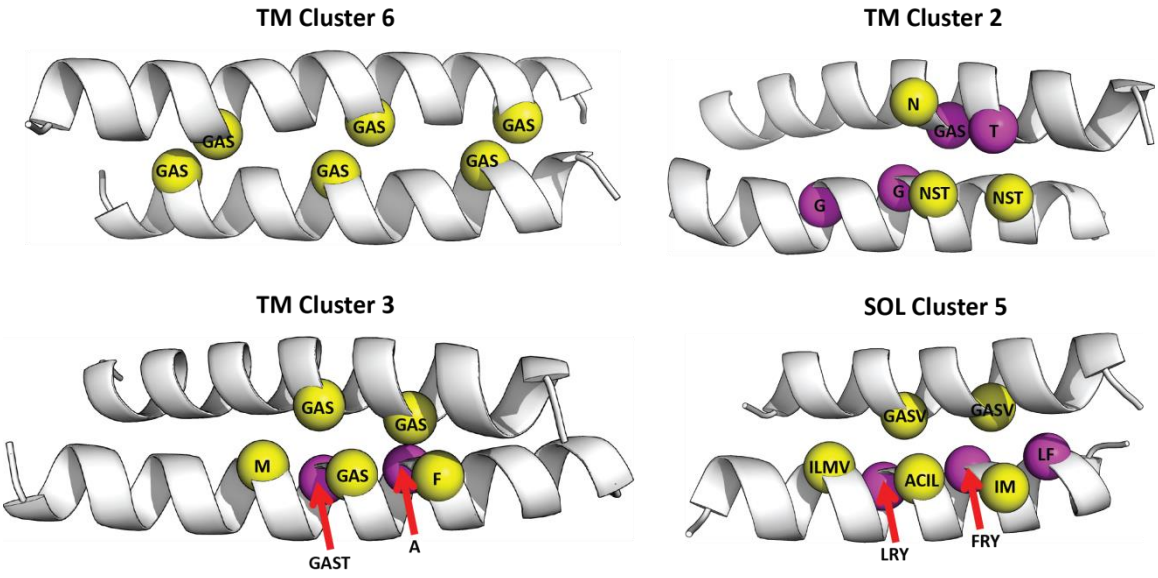


Figure 3. Packing preference of tight clusters. In heptad repeats, positions *a* are colored yellow; in tetrad repeats, positions *a* and *b* are colored yellow and magenta, respectively. Residues are labeled in one-letter representation. Combinations of amino acids mean more than one significant residues.

2.8 Tables

Designation	Category	Cluster No.	No. of members	Percentage*,%	Crossing angle†,°	Distance†, Å
Frequent left-handed						
Antiparallel	TM (WD)	1	130	29.2	-156.5 (10.1)	8.61 (0.89)
	TM	1	221	21.7	-159.9 (7.7)	9.55 (0.94)
	TM	6	86	8.4	-157.0 (6.5)	8.32 (0.56)
	TM	7	63	6.2	-159.4 (9.3)	11.72 (0.76)
Parallel	TM (WD)	4	42	9.4	13.8 (16.6)	9.77 (1.18)
	TM	4	79	7.8	12.2 (10.5)	10.56 (0.76)
Frequent right-handed						
Antiparallel	TM (WD)	2	71	16.0	146.4 (13.6)	8.57 (0.99)
	TM	2	72	7.1	146.6 (10.2)	8.71 (0.85)
	TM (WD)	5	29	6.5	178.0 (20.8)	9.14 (1.47)
	TM	5	81	7.9	165.0 (6.8)	10.44 (0.81)
Parallel	TM (WD)	3	57	12.8	-37.9 (7.5)	7.93 (0.88)
	TM	3	77	7.6	-38.1 (6.9)	8.25 (0.94)

Table 1. Comparison of the top 7 TM Clusters and the corresponding ones in the WD analysis. *Population of the each cluster is based on the quantity of alignments occupied in the total of the 20 clusters. †Values are measured on the most populated 12-residue windows of the clusters in our analysis and standard deviations are shown in parentheses.

Designation	Category	Cluster No.	No. of members	Percentage*,%	Crossing angle†,°	Distance†, Å	RMSD‡, Å
Frequent left-handed							
Antiparallel	TM	1	221	21.7	-159.9 (7.7)	9.55 (0.94)	
	SOL	1	754	27.3	-156.7 (7.0)	9.44 (0.69)	0.59
	TM	6	86	8.4	-157.0 (6.5)	8.32 (0.56)	
	SOL	1	753	27.3	-156.7 (7.0)	9.44 (0.69)	1.20
	TM	7	63	6.2	-159.4 (9.3)	11.72 (0.76)	
	SOL	6	217	7.9	-151.3 (7.6)	11.01 (0.80)	1.30
	TM	7	63	6.2	-159.4 (9.3)	11.72 (0.76)	
	SOL	10	153	5.5	-169.0 (4.6)	11.39 (0.82)	1.24
Parallel	TM	4	79	7.8	12.2 (10.5)	10.56 (0.76)	
	SOL	2	275	10.0	23.0 (6.4)	9.93 (0.83)	0.91
Frequent right-handed							
Antiparallel	TM	2	72	7.1	146.6 (10.2)	8.71 (0.85)	
	SOL	3	263	9.5	136.0 (9.8)	9.48 (0.69)	1.83
	TM	5	81	7.9	165.0 (6.8)	10.44 (0.81)	
	SOL	4	196	7.1	160.1 (5.3)	10.71 (0.61)	0.53
Parallel	TM	3	77	7.6	-38.1 (6.9)	8.25 (0.94)	
	SOL	5	176	6.4	-42.2 (6.4)	8.82 (0.72)	0.65

Table 2. Comparison of the top 7 TM Clusters and their SOL structural counterparts. *Population of the each cluster is based on the quantity of alignments occupied in the total of the 20 clusters. †Values are measured on the most populated 12-residue windows of the clusters in our analysis and standard deviations are shown in parentheses. ‡The values are measured on the 12-residue windows on the centroids with the smallest RMSDs around the most populated 15-residue regions.

2.9 Supplemental Figures

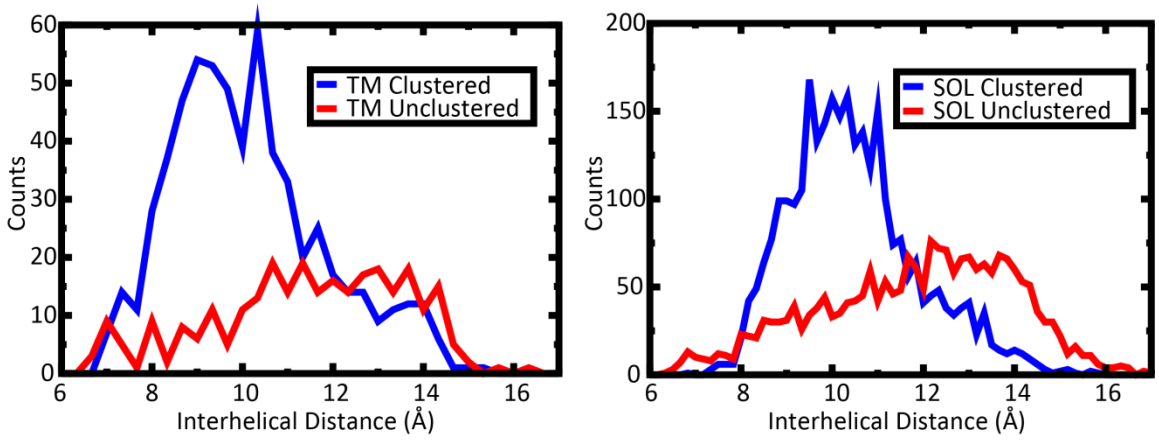
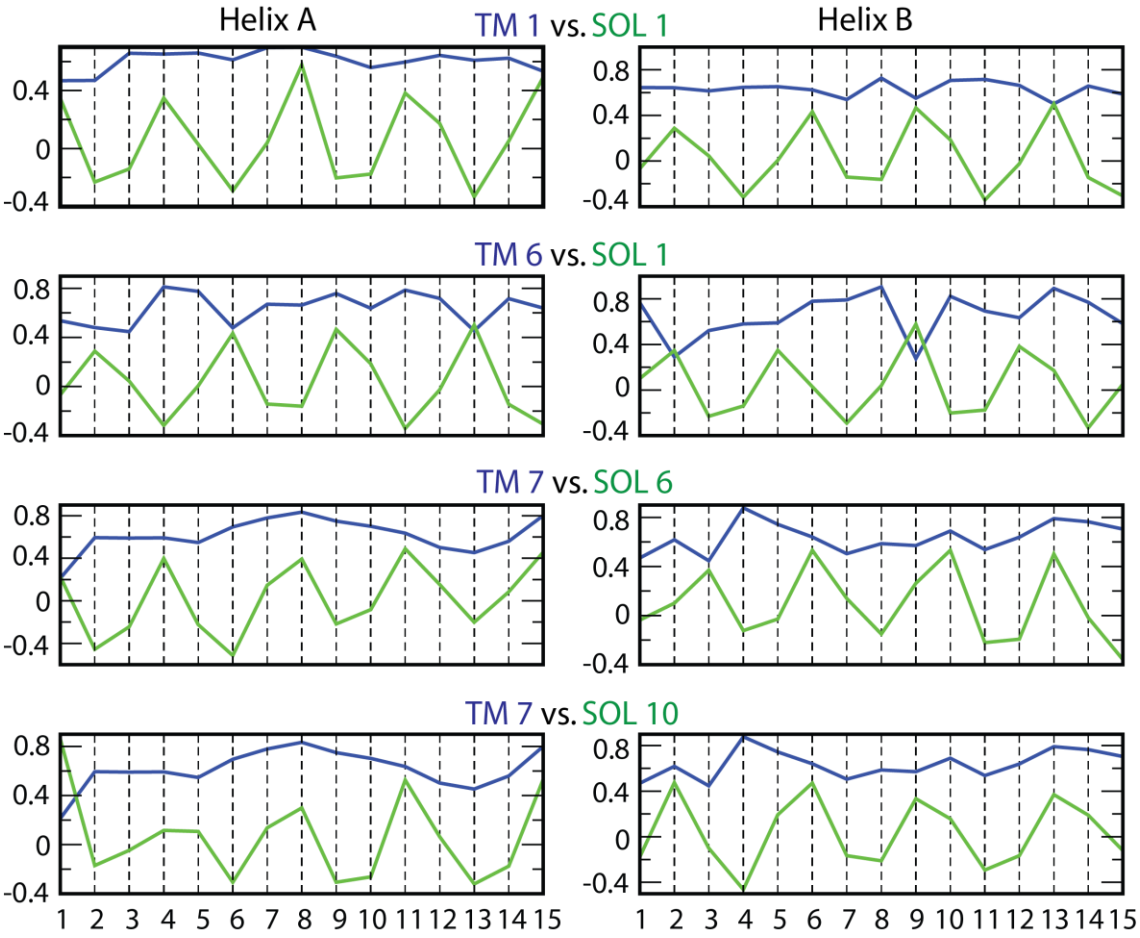


Figure S1. Distribution of interhelical distance of clustered and unclustered helix pairs.



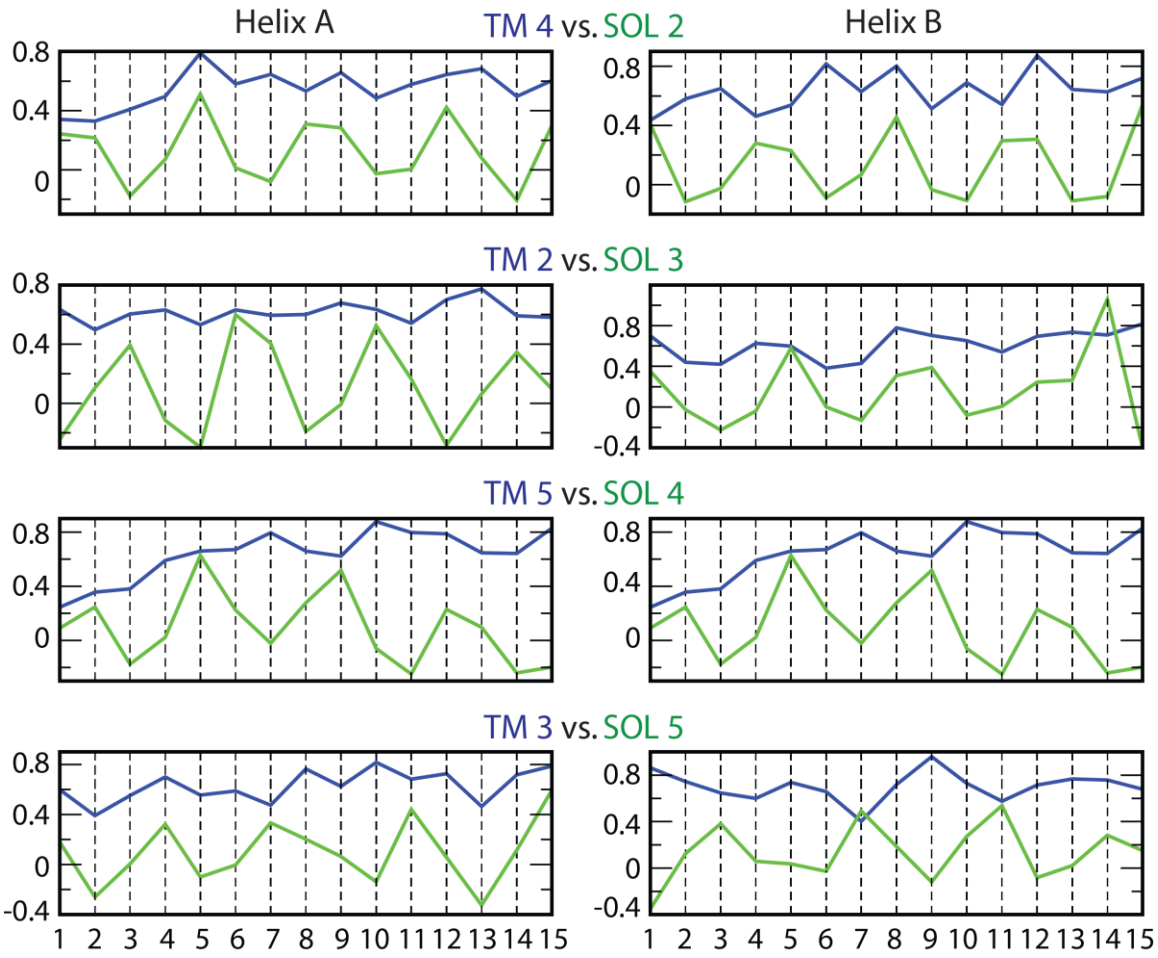


Figure S2. Hydrophobicity of top 7 TM clusters and SOL structural counterparts. Average hydrophobicity was calculated on the windows of 15 most populated positions of top 7 TM clusters. Structurally matched windows from SOL clusters were used to make comparison. Colors of the curves were indicated in the cluster IDs.

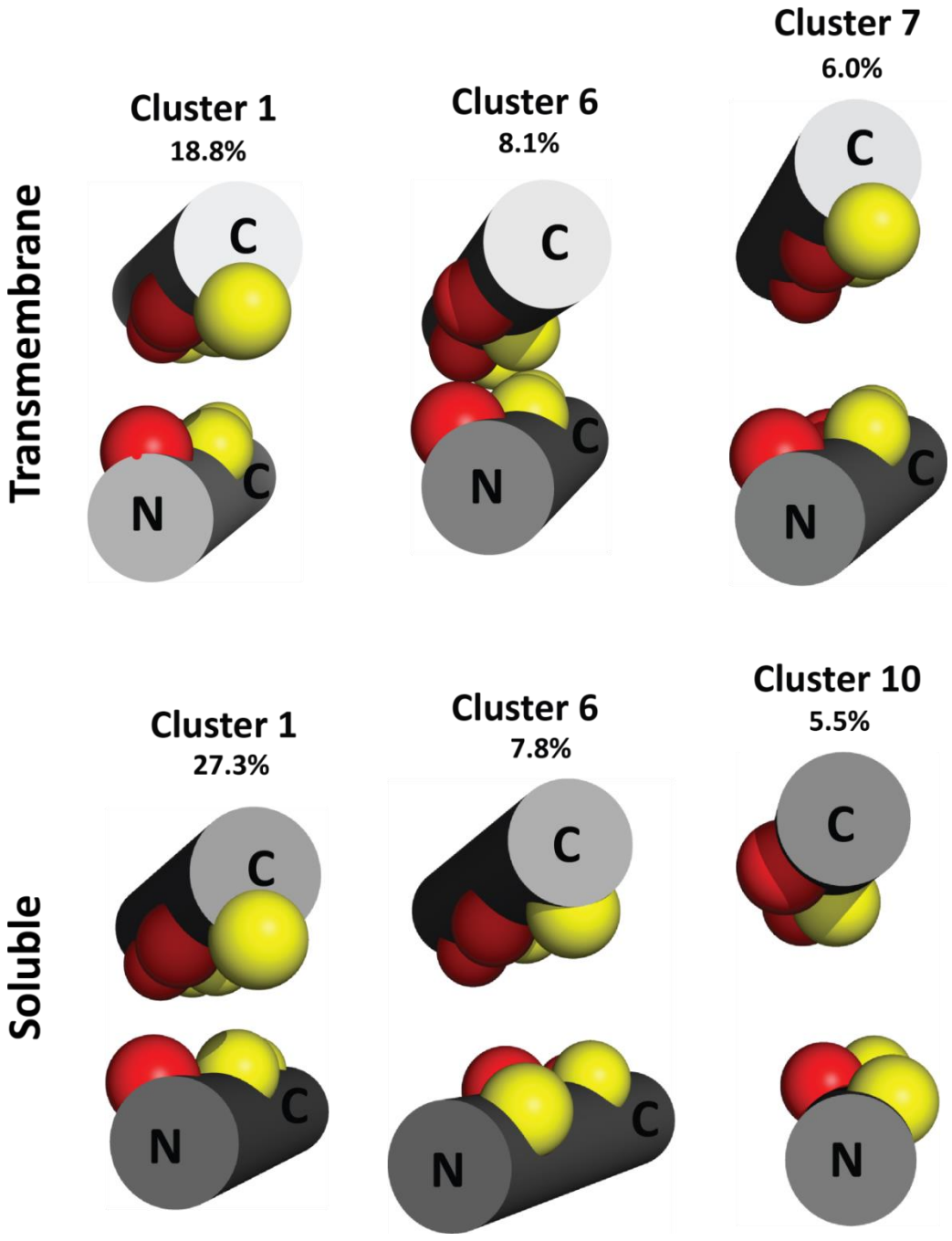


Figure S3. Spatial arrangements of the helices in left-handed antiparallel clusters. The population of the clusters is shown. In heptad repeats, positions *a* and *d* are colored yellow and red, respectively. The N- and C-termini of the helices are labeled.

2.10 Supplemental Tables

Amino acid	TM	SOL
Arg	0.0220	0.0542
Lys	0.0171	0.0602
Asp	0.0124	0.0582
Gln	0.0136	0.0414
Asn	0.0132	0.0400
Glu	0.0177	0.0754
His	0.0220	0.0226
Ser	0.0451	0.0552
Thr	0.0559	0.0507
Pro	0.0256	0.0405
Tyr	0.0337	0.0345
Cys	0.0109	0.011
Gly	0.0839	0.0663
Ala	0.1165	0.0911
Met	0.0407	0.0215
Trp	0.0260	0.0127
Leu	0.1624	0.1006
Val	0.0992	0.0670
Phe	0.0840	0.0385
Ile	0.0981	0.0582

Table S1. Background distributions of amino acids in transmembrane and soluble proteins.

2.11 References

1. Martin, J., et al., *Protein secondary structure assignment revisited: a detailed analysis of different assignment methods.* BMC Struct Biol, 2005. **5**: p. 17.
2. Greene, L.H., et al., *The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution.* Nucleic acids research, 2007. **35**: p. D291-7.
3. Deisenhofer, J., et al., *X-ray structure analysis of a membrane protein complex. Electron density map at 3 A resolution and a model of the chromophores of the photosynthetic reaction center from Rhodopseudomonas viridis.* J Mol Biol, 1984. **180**(2): p. 385-98.
4. Bowie, J.U., *Solving the membrane protein folding problem.* Nature, 2005. **438**(7068): p. 581-9.
5. Wallin, E. and G. von Heijne, *Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms.* Protein Sci, 1998. **7**(4): p. 1029-38.
6. Shai, Y., *Molecular recognition within the membrane milieu: implications for the structure and function of membrane Pproteins.* J Membr Biol, 2001. **182**(2): p. 91-104.
7. White, S.H., *Biophysical dissection of membrane proteins.* Nature, 2009. **459**(7245): p. 344-6.
8. Gimpelev, M., et al., *Helical packing patterns in membrane and soluble proteins.* Biophys J, 2004. **87**(6): p. 4075-86.
9. Walters, R.F. and W.F. DeGrado, *Helix-packing motifs in membrane proteins.* Proc Natl Acad Sci U S A, 2006. **103**(37): p. 13658-63.
10. Lemmon, M.A., et al., *Glycophorin A dimerization is driven by specific interactions between transmembrane alpha-helices.* J Biol Chem, 1992. **267**(11): p. 7683-9.
11. MacKenzie, K.R., J.H. Prestegard, and D.M. Engelman, *A transmembrane helix dimer: structure and implications.* Science, 1997. **276**(5309): p. 131-3.
12. Lemmon, M.A., et al., *A dimerization motif for transmembrane alpha-helices.* Nat Struct Biol, 1994. **1**(3): p. 157-63.
13. Senes, A., M. Gerstein, and D.M. Engelman, *Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions.* J Mol Biol, 2000. **296**(3): p. 921-36.
14. Russ, W.P. and D.M. Engelman, *The GxxxG motif: a framework for transmembrane helix-helix association.* J Mol Biol, 2000. **296**(3): p. 911-9.
15. Gernert, K.M., et al., *The Alacoil: a very tight, antiparallel coiled-coil of helices.* Protein Sci, 1995. **4**(11): p. 2252-60.
16. Liu, Y., D.M. Engelman, and M. Gerstein, *Genomic analysis of membrane protein families: abundance and conserved motifs.* Genome Biol, 2002. **3**(10): p. research0054.
17. Unterreitmeier, S., et al., *Phenylalanine promotes interaction of transmembrane domains via GxxxG motifs.* J Mol Biol, 2007. **374**(3): p. 705-18.

18. Herrmann, J.R., et al., *Complex patterns of histidine, hydroxylated amino acids and the GxxxG motif mediate high-affinity transmembrane domain interactions.* J Mol Biol, 2009. **385**(3): p. 912-23.
19. Adamian, L. and J. Liang, *Interhelical hydrogen bonds and spatial motifs in membrane proteins: polar clamps and serine zippers.* Proteins, 2002. **47**(2): p. 209-18.
20. Liang, J., *Experimental and computational studies of determinants of membrane-protein folding.* Curr Opin Chem Biol, 2002. **6**(6): p. 878-84.
21. Sal-Man, N., et al., *Specificity in transmembrane helix-helix interactions mediated by aromatic residues.* J Biol Chem, 2007. **282**(27): p. 19753-61.
22. Langosch, D. and I.T. Arkin, *Interaction and conformational dynamics of membrane-spanning protein helices.* Protein Sci, 2009. **18**(7): p. 1343-58.
23. Hedin, L.E., K. Illergard, and A. Elofsson, *An introduction to membrane proteins.* J Proteome Res, 2011. **10**(8): p. 3324-31.
24. Varriale, S., et al., *An evolutionary conserved motif is responsible for immunoglobulin heavy chain packing in the B cell membrane.* Mol Phylogenet Evol, 2010. **57**(3): p. 1238-44.
25. Gratkowski, H., et al., *Cooperativity and specificity of association of a designed transmembrane peptide.* Biophys J, 2002. **83**(3): p. 1613-9.
26. Zhou, F.X., et al., *Polar residues drive association of polyleucine transmembrane helices.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**: p. 2250-5.
27. Lawrie, C.M., E.S. Sulistijo, and K.R. MacKenzie, *Intermonomer hydrogen bonds enhance GxxxG-driven dimerization of the BNIP3 transmembrane domain: roles for sequence context in helix-helix association in membranes.* Journal of molecular biology, 2010. **396**: p. 924-36.
28. Han, Q., et al., *Conserved GXXXG- and S/T-like motifs in the transmembrane domains of NS4B protein are required for hepatitis C virus replication.* Journal of virology, 2011. **85**: p. 6464-79.
29. Wei, P., et al., *The dimerization interface of the glycoprotein Ibβ transmembrane domain corresponds to polar residues within a leucine zipper motif.* Protein science : a publication of the Protein Society, 2011. **20**: p. 1814-23.
30. White, S.H. and W.C. Wimley, *Membrane protein folding and stability: physical principles.* Annu Rev Biophys Biomol Struct, 1999. **28**: p. 319-65.
31. Grigoryan, G. and W.F. Degrado, *Probing designability via a generalized model of helical bundle geometry.* J Mol Biol, 2011. **405**(4): p. 1079-100.
32. Choma, C., et al., *Asparagine-mediated self-association of a model transmembrane helix.* Nat Struct Biol, 2000. **7**(2): p. 161-6.
33. Meindl-Beinker, N.M., et al., *Asn- and Asp-mediated interactions between transmembrane helices during translocon-mediated membrane protein assembly.* EMBO Rep, 2006. **7**(11): p. 1111-6.
34. Senes, A., D.E. Engel, and W.F. DeGrado, *Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs.* Curr Opin Struct Biol, 2004. **14**(4): p. 465-79.
35. Kleiger, G., et al., *GXXXG and AXXXA: common alpha-helical interaction motifs in proteins, particularly in extremophiles.* Biochemistry, 2002. **41**(19): p. 5990-7.

36. Benjamini, A. and B. Smit, *Robust driving forces for transmembrane helix packing.* Biophys J, 2012. **103**(6): p. 1227-35.
37. Lomize, M.A., et al., *OPM database and PPM web server: resources for positioning of proteins in membranes.* Nucleic Acids Res, 2012. **40**(Database issue): p. D370-6.
38. Wang, G. and R.L. Dunbrack, Jr., *PISCES: a protein sequence culling server.* Bioinformatics, 2003. **19**(12): p. 1589-91.
39. Tusnady, G.E., Z. Dosztanyi, and I. Simon, *PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank.* Nucleic Acids Res, 2005. **33**(Database issue): p. D275-8.
40. Zhang, J. and G. Grigoryan, *Mining tertiary structural motifs for assessment of designability.* Methods Enzymol, 2013. **523**: p. 21-40.
41. Bansal, M., S. Kumar, and R. Velavan, *HELANAL: a program to characterize helix geometry in proteins.* J Biomol Struct Dyn, 2000. **17**(5): p. 811-9.
42. Eisenberg, D., et al., *Analysis of membrane and surface protein sequences with the hydrophobic moment plot.* J Mol Biol, 1984. **179**(1): p. 125-42.

Chapter 3

Stability of a Peptide Designed for Selective Carbon Nanotube Hybridization

3.1 Overview

Biological polymers hybridized with single-walled carbon nanotubes (SWCNTs) have elicited much interest recently for applications in SWCNT-based sorting as well as biomedical imaging, sensing, and drug delivery. Recently, de novo designed peptides forming a coiled-coil structure have been engineered to selectively disperse SWCNT of a certain diameter. Here we report on a study of the binding strength and structural stability of the hybrid between such a "HexCoil-Ala" peptide and the (6,5)-SWCNT. Using the competitive binding of a surfactant, we find that affinity strength of the peptide ranks in comparison to that of two single-stranded DNA sequences as $(GT)_{30}$-DNA > HexCoil-Ala > $(TAT)_4T$-DNA. Further, using replica exchange molecular dynamics (REMD), we show that multiple anti-parallel HexCoil-Ala strands are needed for stability on the (6,5)-SWCNT; configurations of one or two strands become disordered. Detailed analysis of the simulation results showed similarities and differences from the original design. While one of two distinct helix-helix interfaces of the original model was largely retained, a second interface showed much greater variability. These conformational differences allowed an aromatic tyrosine residue designed to lie along the solvent-exposed surface of the protein instead to penetrate between the two helices and directly contact the SWCNT. These insights will inform future designs of SWCNT-interacting peptides.

3.2 Introduction

Much effort has been expended in recent years studying and developing desirable properties and applications of the single-walled carbon nanotube (SWCNT). These include their ability as strengthening agents for composite materials [1], construction of field-effect transistor devices [2, 3], and *in vitro/in vivo* imaging and targeted delivery agents in biomedical applications [4-8]. As objects foreign to cells, SWCNTs present a certain degree of cytotoxicity [9, 10]. However, this can be reduced greatly by appropriate surface functionalization [11-13]. Additionally, upon production, SWCNTs tend to clump together in bundles of mixed chirality (electronic species) due to their high aspect ratios and hydrophobic surface[14, 15]. Numerous recent methods have been developed to solubilize and sort SWCNTs by length [16, 17], diameter [18], and electronic structure by hybridization with a dispersant molecule [19]. The dispersant molecule can range from small inorganic surfactants (e.g., sodium dodecyl sulfate) [20] to biological polymers (short DNA oligomers or peptides) [21, 22]. The ability of certain short strands of DNA to recognize particular SWCNTs from a chirality-diverse mixture, enabling single-species purification, has also been demonstrated [23].

The design of peptides for SWCNT dispersion has also been investigated [21, 24-26]. In general, a peptide with sufficient hydrophobic residues located at appropriate sites along its backbone will be able to disperse a SWCNT in aqueous medium to some extent. By designing peptide sequences to promote the arrangement of hydrophobic residues to one side of an alpha helix, SWCNT dispersion abilities were shown to be significantly increased [21, 24]. More recent studies have attempted to selectively disperse SWCNTs

of a particular diameter or chirality from a mixture using designed peptides [27]. Grigoryan et al. have developed a *de novo* method of peptide design using sequences known to form α helices that then assemble into hexa-coiled supramolecular structures [27]. By controlling the diameter of the hexa-coiled structure through sequence modulation, they have been able to selectively disperse (6,5) and (8,3)-SWCNTs from mixtures. When design is based on the primary structure, it is implied that the peptide will assume some adsorbed conformation likely different from its solution state. When stable secondary structures are designed, as in the example just cited, it is assumed that this structure will unravel by virtue of interaction with the SWCNT, which may or may not be the case [28].

Here, we study the affinity of a particular 30-amino acid long peptide, "HexCoil-Ala", for the (6,5)-SWCNT through experimentation and simulation [27]. This alanine-rich sequence has been shown to singly-disperse SWCNTs, as indicated by strong near-infrared (NIR) photoluminescence [20]. We used surfactant-induced displacement of adsorbed molecules from the SWCNT surface to rank and quantify binding strength compared to chosen DNA sequences [29]. Ranking was then confirmed by creating dispersions of raw SWCNTs in mixtures of peptide or DNA and surfactant. NIR absorbance measurements were used to identify which type of molecule remained on the SWCNT. Using replica exchange molecular dynamics (REMD) simulation, we probed how the stability of the peptide-SWCNT structure depends on the number of peptide molecules adsorbed on the SWCNT surface. The symmetry of the original hexamer dictated two distinct helix-helix interfaces that were considered in the design process. Simulations showed that only one of the two interfaces remained stable on the 50 ns

REMD time scale, and rearrangements were observed in the interaction of the helices, specifically that of an aromatic Tyr residue with the SWCNT due to π–π interactions.

3.3 Methodologies

As described by Grigoryan et al. [27], dispersions of HexCoil-Ala peptide were created using Comocat nanotubes (*SWeNT*). First, 1 mg of previously synthesized, purified, and lyophilized HexCoil-Ala (AEAESALEYAQQALEKAQLALQAARQALKA) was added to 0.1 mg of raw nanotubes in a 100 mM phosphate buffer at pH 7.4. The solution was then probe-sonicated (*Branson*) at 8 Watts for 90 minutes in an ice-cooled bath followed by 6 hours of centrifugation (*Eppendorf*) at 16,000 times the force of gravity. The resultant supernatant was then extracted and used for analysis. Additionally, hybrids of DNA sequences $(GT)_{30}$ or $(TAT)_4T$, and Comocat nanotubes, in a weight ratio of 1:1, were also created using the same procedure for comparison with peptide-SWCNT.

Initial absorbance and fluorescence spectra of the peptide-SWCNT dispersion were measured. A UV/Vis/NIR spectrophotometer (*Varian Cary50*) was used to measure the absorbance spectrum from 200-1100 nm of the dispersion in a quartz microcuvette. A prominent NIR peak was observed at 992 nm, indicative of the E_{11} bandgap transition for the (6,5)-SWCNT. Furthermore, a two-dimensional excitation/emission NIR fluorescence map (*Horiba Yvon Jobin Fluorolog-3*) of the peptide-SWCNT dispersion was measured. The excitation and emission ranges were 500-800 nm and 900-1200 nm, respectively, with a slit width of 8 nm and data interval of 3 nm. Again, the dominant peak corresponded to a (6,5)-SWCNT with excitation/emission pair of 569/992 nm.

In accordance with a previously used method [29], a small-molecule surfactant, sodium dodecylbenzene sulfonate (SDBS), was used in an attempt to displace the peptide off the surface of a (6,5)-SWCNT. A solution of 0.2 wt % SDBS in the same 100 mM phosphate buffer used for SWCNT dispersion was held at 60^{o} C in the quartz cuvette. In a 1:1 v/v ratio, peptide-SWCNT solution was introduced into the cuvette and pipette-mixed at time zero. The effective SDBS concentration was reduced to 0.1 wt %, less than the critical micelle concentration (CMC) of the surfactant [29]. Over the course of the next 30 minutes, the NIR absorbance was scanned from 950-1050 nm in one minute intervals to monitor the progress of the surfactant displacement reaction. The procedure was then repeated using DNA with sequences $(GT)_{30}$ and $(TAT)_4T$-SWCNT. In addition, in place of SDBS, a different surfactant, sodium cholate, was used to attempt surfactant exchange. Displacement by the surfactant causes a solvatochromic shift in the peak of the absorbance spectrum. By tracking this shift the relative progress and speed of the reaction can be monitored.

Binary dispersions (mixtures of SDBS and peptide, or SDBS and $(GT)_{30}/(TAT)_4T$, in equal mass ratios) were created. The raw SWCNT sample was then sonicated in the presence of this mixture of molecules (10:10:1 by weight) for 90 minutes and centrifuged as previously described, allowing the surfactant and peptide or DNA to compete for the SWCNT surface. The supernatants' NIR absorbance spectra were measured following this procedure.

In addition to changes in the absorbance, a final fluorescence map of the peptide-SWCNT solution was measured after surfactant exchange using the same parameters as described

previously. Circular dichroism (CD) in the far-UV (190-240 nm) was measured (*Jasco J-815*) using a quartz cuvette with a path length of 1 mm to investigate characteristics of the secondary structure of the peptide-SWCNT hybrids as they encounter surfactant. The surfactant chosen for this study was sodium dodecyl sulfate (SDS) for its relatively low absorbance in the UV region as compared to SDBS or sodium cholate. The peptide remained at 1 mg/mL with SDS at a concentration of 0.1 % wt.

For the MD study, we began by using the HexCoil-Ala structure available on the RCSB protein data bank (PDB) as structure – 3S0R. This file contains two chains, A and B, in an antiparallel configuration. In the first simulation, one strand of the HexCoil-Ala was placed on a (6,5)-SWCNT in an orientation permitting hydrophobic peptide residues to be in close proximity with the SWCNT surface (Figure 1a). The SWCNT was 8.12 nm long with a diameter of 0.746 nm. The length was chosen such that one end of the frozen SWCNT would exactly adjoin its periodic image thus creating an infinitely long SWCNT. The peptide-SWCNT hybrid was then solvated in an $8.12 \times 5.00 \times 5.00$ nm water-box containing approximately 6,200 TIP3P model [30] water molecules with the appropriate number of sodium counter-ions to balance the net-negatively charged peptide (Figure 1b). Periodic boundary conditions were applied in all directions with long-range electrostatics interactions calculated using the particle mesh Ewald method [31]. All structures were visualized in *VMD* [32].

The method of replica exchange MD (REMD) accesses a greater fraction of available microstates by overcoming high energy barriers [33, 34], and has been used in the past to determine equilibrium structures in simulations of DNA-SWCNT hybrids [35-37]. Here,

the Gromacs 4.5.3 simulation package [38-40] was used in conjunction with the Amber03d [41] force field for REMD simulation. Forty replicas were simulated in parallel with temperatures ranging from 296 K to 587 K. The replica temperatures were chosen such that exchange acceptance ratios between the replicas remained around 10% with an exchange time of 1 ps. The single strand peptide-SWCNT simulation was then run for 200 ns of REMD, for a total computation time of 40×200 ns = 8 μs. The time step of the simulation was 2 fs. Clustering was then performed on the last 100 ns of the 300 K trajectory using the backbone peptide atoms constrained to a root mean squared deviation (RMSD) of 0.3 nm.

The two largest clusters from the single-strand peptide-SWCNT configuration, representing 5% and 4% of the trajectory, respectively, were used to create the initial structure for the two-strand simulation. The new structure was then re-equilibrated with water and counter-ions. Again, REMD was performed on this configuration for 200 ns at the same 40 temperatures, and backbone clustering was performed on the last 100 ns of the 300 K trajectory.

In the case of the 6 strand, hexa-coiled peptide-SWCNT configuration, three copies of the PDB file (3S0R) were placed around the exterior of the same (6,5)-SWCNT. The structure was again re-equilibrated with water and counter-ions and run for 50 ns of REMD simulation. Analysis was performed on the final 45 ns of data, allowing for 5 ns of equilibration. Reported interhelical distances were calculated from the center 10 residues of each peptide chain.

We defined helical structure in terms of which regions of the Ramachandran map were occupied. The α_h region of the (φ , ψ) map was defined as $\varphi \in [-100°, -30°]$ and $\psi \in [-67°, -7°]$. Residues which lay within the α_h region of the Ramachandran map were denoted as helical (h). All residues outside the α_h region were defined as "coil" (c). A helical segment was one which had at least three consecutive residues whose (φ , ψ) angles fall within the α_h boundaries (i.e., the smallest helix is ...chhhc...). The fraction of helix for a given residue in the simulation was calculated as the fraction of time spent by that residue within helical segments.

3.4 Results and Discussion

To probe the structural integrity of the synthesized peptide-SWCNT complex, several binding affinity experiments were performed. First, using the method of surfactant-induced displacement (exchange), a relative measure of the hybrid's stability was determined. It is known that surfactant SDBS has a higher affinity for the surface of a SWCNT than short strands of DNA and thereby displaces the latter at a characteristic rate [29]. Surfactant exchange of the peptide, monitored through changes in NIR absorbance of the (6,5)-SWCNT, was attempted with SDBS as well as another surfactant, sodium cholate. Figures 2a,b, show that the effect of the two different surfactants on the peptide-SWCNT hybrid is almost negligible at the elevated temperature of 60^{o}C over the course of 10 minutes. The effect of SDBS on the peptide-SWCNT sample is to broaden the peak with a significant appearance of a blue-shifted shoulder at 978 nm, characteristic of an SDBS-covered (6,5)-SWCNT, and an accompanying slight decrease of absorbance at 990 nm. Negligible effect of surfactant is seen in circular dichroism data of the peptide

whether it is on the SWCNT or off it (see supplemental information, section S1). Figures 2a and b should be compared to Figures 2c and d, which show the change in the NIR spectrum due to displacement by SDBS of the DNA sequence $(GT)_{30}$ or $(TAT)_4T$. These DNA sequences have been chosen for their, respectively, strong and weak binding affinities to the (6,5)-SWCNT.[26, 29] The $(GT)_{30}$ sample shows very little change over the course of the reaction; particularly absent is the blue-shifted shoulder at 978 nm. In contrast, $(TAT)_4T$ is almost immediately displaced from the SWCNT surface, evident in the blue-shifted peak. In addition, DNA-SWCNT surfactant exchange experiments show the existence of the (7,5)-SWCNT in the dispersion, with a starting absorbance peaking at 1040 nm. Note its absence in the peptide-SWCNT spectra, indicating the peptide's preferential ability to disperse the smaller-diameter nanotube, (6,5). Strong and preferential binding of the peptide to the (6,5)-SWCNT, relative to DNA sequence $(TAT)_4T$, was confirmed by two-dimensional fluorescence maps of the 'ending' samples in Figure 2a and d (see supplemental information, section S1).

Previous work on displacement of DNA molecules by SDBS from an SWCNT has shown that the process occurs by an initial fast step which was interpreted as conversion of SWCNTs with pre-existing defects to coating by SDBS [29]. This is followed by a slower second step with rate of displacement by SDBS limited presumably by the nucleation of defects. On this basis, we suggest that the emergent shoulder in Figure 2a represents displacement by SDBS of those hexa-coiled peptides that have some form of defect. Over the time-frame of the experiment, it is clear that the remaining majority of SWCNTs in the sample strongly resist displacement by SDBS. By qualitatively

comparing rates of SDBS exchange, we can rank the affinity of the examined biopolymers to the (6,5)-SWCNT as $(GT)_{30}$ > HexCoil-Ala > $(TAT)_4T$.

As another test of their binding affinities for the (6,5)-SWCNT, binary dispersions were created. The absorbance spectra for dispersions of $(GT)_{30}$, $(TAT)_4T$, or HexCoil-Ala mixed with SDBS and SWCNT are shown in Figure 2e. Observe that the HexCoil-Ala-SDBS-SWCNT absorbance spectrum has a peak at 992 nm but with a significant blue-shifted shoulder. Consistent with surfactant exchange data in Figure 2a, this suggests that two stable species exist in solution; surfactant-covered and peptide-covered SWCNT. By comparison, the $(GT)_{30}$ sequence out-competes SDBS for coverage of the SWCNT surface, as indicated by the fact that the absorbance peak remains centered at 992 nm. In contrast, the $(TAT)_4T$ sequence is out-competed by SDBS since the absorbance peak shifts to 980 nm. These experiments confirm that the relative binding strengths to the (6,5)-SWCNT can be ranked as $(GT)_{30}$ > HexCoil-Ala > $(TAT)_4T$.

The HexCoil-Ala-(6,5)-SWCNT hybrid structure has additionally been investigated by using REMD molecular simulation. We obtained what can be regarded as representative equilibrium structures for one, two, and six strands of HexCoil-Ala peptide, and found very significant differences in structure among these three cases. Figure 3a shows that a single strand of HexCoil-Ala loses the large majority of its alpha-helical nature, unwrapping on the surface of the (6,5)-SWCNT. From a clustering analysis, found in the methods section, the top two clusters only represented 5% and 4% of the total population, respectively. The two-strand configuration shows similar behavior, Figure 3b. That is,

the equilibrium structures of both strands are in a disordered state and interactions between adjacent peptide strands appear to be minimal.

By sharp contrast, for the case of six strands of the HexCoil-Ala peptide, i.e., for the reported canonical form [27], the hexacoiled structure remains stable over the time frame of the REMD simulation, Figure 3c. The six strands, situated in anti-parallel configuration, remain adsorbed to the surface of the SWCNT in alpha-helical arrangements. In Figure 3d, the fraction of time that each residue is in a helical state is plotted for the one, two, and six HexCoil-Ala strand configurations. The one and two strand plots confirm that much of the alpha-helical structure has been lost. In contrast, with some variation, the six strand configuration retains the majority of its helicity over the course of the simulation. Convergence data can be found in supplemental information, section S3. Additionally, residues near the ends of the alpha helix exhibit a certain degree of disorder as shown by a drop in fraction helix. On a long SWCNT with multiple strands along the nanotube length, this loss of structure will likely be quenched by additional hexacoiled structures placed on either side of the one in question.

The homohexamer HexCoil-Ala shows inhomogeneous patterns of helix-helix association upon binding to SWCNT. In the simulations, the overall configurations show little geometric variability for the hexamer (Figure S4a) after the initial “equilibration” step. There were two types of interface designed for HexCoil-Ala: leucine zipper and Ala-Coil (Figure S5). The former is a well-studied structural motif in proteins, while the latter is a tight antiparallel coiled-coil motif [42] that is common in transmembrane proteins [43], and occurs more rarely in water-soluble proteins according to Chapter 1.

The leucine zipper interface is formed along the interfaces between Chains B+C, D+E and F+G, and the Ala-Coil interface by Chains C+D, E+F and G+B (Figure 4a). The Ala-Coil interface displays much greater structural variability among different chains in its interhelical distance (Figure 4c, 4d). The leucine zipper interface has a trimodal distribution in interhelical distances among the chain pairs (Figure 4c), ranging from 9.5 to 11.0 Å. These values are well within the range seen for typical antiparallel helix dimers with Leu residues at similar positions in the sequence [44].

The plasticity of the Ala-Coil interface causes large deviations from the designed model. In the tetrameric HexCoil-Ala crystal structure solved in the absence of SWCNTs (PDB ID: 3S0R), the Ala-Coil interface adopts an interhelical distance as small as 8.55 Å, quite similar to that intended in the original model of the hexamer. In the hexamer model designed to wrap SWCNT, the distance is 8.67 Å, but in the simulations, the average distance increases quite substantially to 13.0 Å, 13.0 Å and 9.86 Å, for Chains C+D, E+F and G+B, respectively. In contrast to the crystal structure and the original design of the hexmaer, the alanine residues form much fewer interchain contacts in the simulations (Figure 4b). In comparison, the configuration of the Leu-Zipper interface is robust. The interhelical distance in the tetramer crystal structure and the hexamer model is 10.5 Å and 10.6 Å, respectively. The average distance of Chains B+C, D+E and F+G is 10.3 Å, 11.2 Å and 10.4 Å, respectively. Interestingly, the great interhelical distance of the Ala-Coil interface extrudes Chain F from SWCNT (Figure 4a), causing helices E, F, and G to form three fourths of classical four-helix bundle geometry. Thus the local arrangement of Chains E, F and G is reminiscent of the tetramer crystal structure (Figure S4b).

There is a single aromatic tyrosine residue in each monomer of the homohexamer. Originally, the Tyr residue was included as a spectroscopic label, and positioned at the helix-helix interface of the Ala-Coil directed outward towards solvent. However, the strong affinity of aryl groups for SWCNTs becomes apparent in the simulations, and might contribute to the deviation from the original, highly symmetric bundle geometry. There are three main clusters in the space of the distance from the phenyl ring atoms on the Tyr residue to SWCNT and the interhelical distance (Figure 4d). In Chains C+D, phenol groups always contact the SWCNT. Chains E+F and G+B both have two configurations: one phenol group is pointing inward and the other is tipping outside, and both of them are directed outward. However, Chains E+F occupy two separate main clusters with distinct interhelical distances, while Chains G+B have a continuous distribution of interhelical distance. The effect is to introduce a wider gap between the helices when the Tyr residues are able to penetrate into direct contact with the SWCNT. Thus the three chain pairs have distinct configurations for helix-helix interaction both in the leucine zipper and Ala-Coil interfaces.

Based on this analysis we can now speculate on the deviation of the observed structure from the design. The leucine zipper motif is greatly stabilized in water by the interdigitation of large apolar Leu side chains. By contrast, the smaller hydrophobic driving force for burial of the Ala residues at the Ala-coil interface makes this structure less stable. Indeed, the Ala-Coil is less frequently observed than the antiparallel leucine zipper in the crystal structures of water soluble proteins.[44] The lower stability of the Ala-Coil motif might provide greater malleability, and allow penetration of the phenol side

chain of Tyr to the SWCNT. Clearly, this is a possibility that should be addressed in future designs.

3.5 Conclusions

We have examined the stability of de novo designed HexCoil-Ala-SWCNT hybrids by means of a surfactant-induced displacement reaction and by dispersion efficiencies in binary mixtures. These methods of ranking can be translated to a variety of non-covalent CNT-wrapping polymers and small molecules. We find that the peptide binds stronger to the (6,5)-SWCNT than DNA sequence $(TAT)_4T$, but weaker than sequence $(GT)_{30}$. Results of REMD molecular simulation approaching equilibrium suggest that the proposed hexacoiled structure is stable relative to configurations containing only one or two strands. The analysis of the hexamer configurations sheds light on the structure of the existing peptide and provides insights for the future design of more specific structures.

3.6. Acknowledgements

Daniel Roxbury, Jeetain Mittal, and Anand Jagota are co-authors of this chapter. This work was supported by the National Science Foundation through grant CMMI-1014960, and a Faculty Innovation Grant (FIG) to Anand Jagota from Lehigh University. This research was also supported in part by the National Science Foundation through TeraGrid resources provided by the Texas Advanced Computing Center (TACC) under grant number [TG-MCB100049]. Support from NIH grant R37GM54616 and from the MRSEC program of NSF to University of Pennsylvania are acknowledged.

3.7 Figures

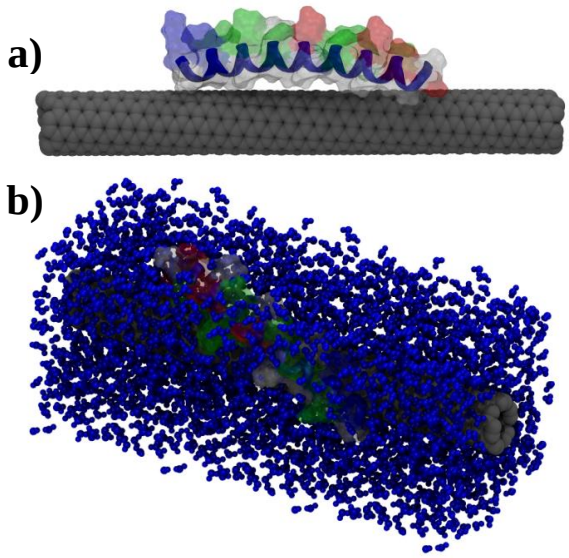


Figure 1. (a) Initial configuration for one strand of HexCoil-Ala peptide placed on a (6,5)-SWCNT and (b) solvated with sodium counter-ions and TIP3P explicit water molecules.

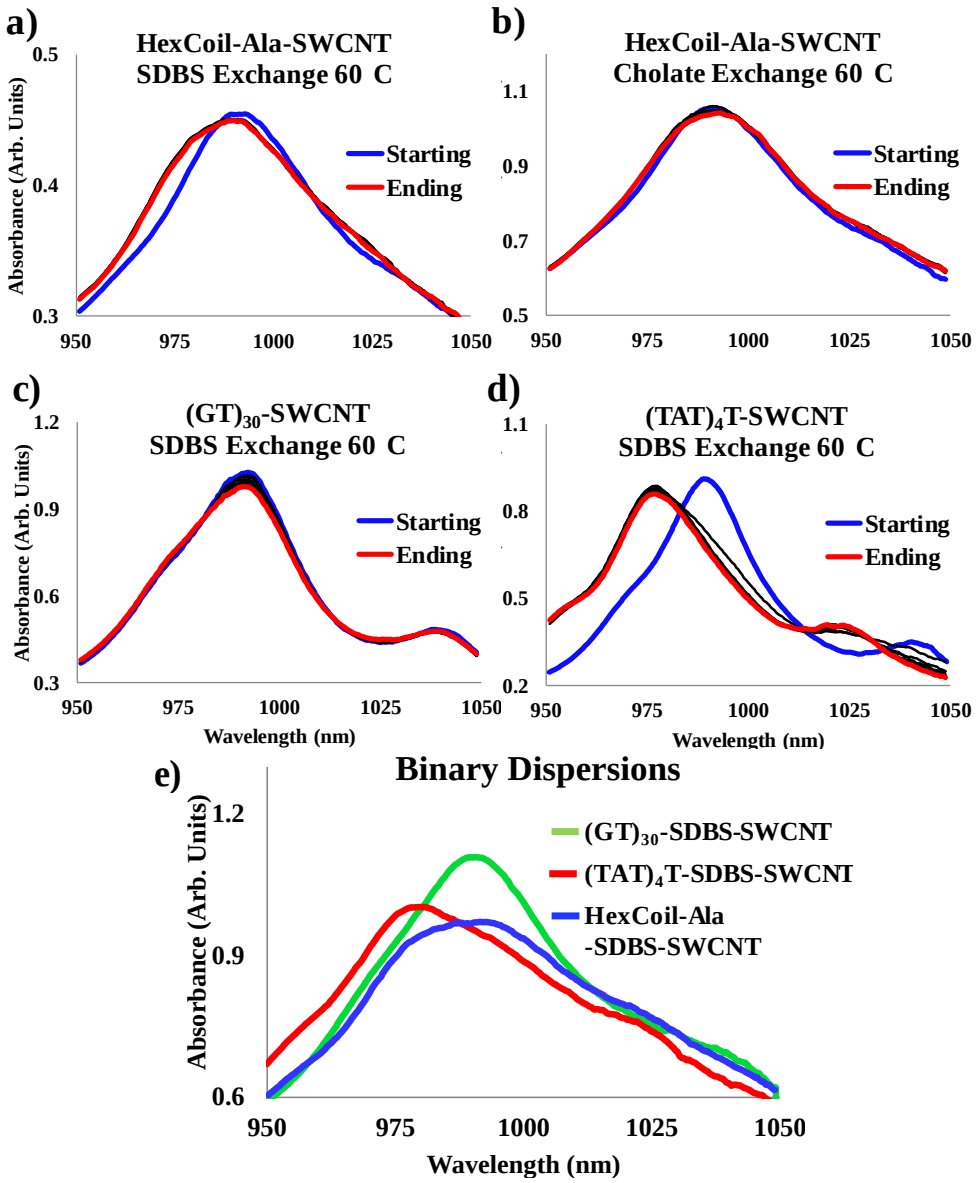


Figure 2. Surfactant exchange performed on a HexCoil-Ala-SWCNT sample subjected to an excess of (a) SDBS and (b) sodium cholate at 60°C for 10 minutes of incubation. For comparison, SDBS exchange is performed on samples of (c) $(GT)_{30}$-SWCNT or (d) $(TAT)_4T$-SWCNT under the same conditions. (e) Binary dispersions of SDBS with $(GT)_{30}$, $(TAT)_4T$, or HexCoil-Ala. The peak position gives an indication of the relative

binding strength of the biopolymer; a blue shift represents replacement by the surfactant molecule.

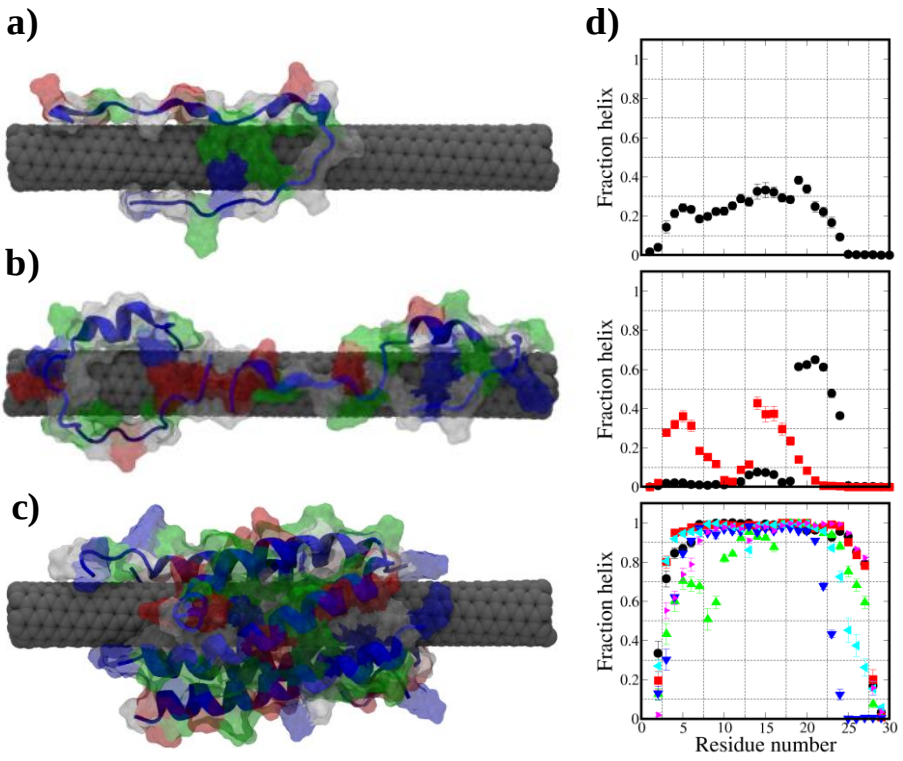


Figure 3. Equilibrium representations of the dominant structure for (a) one, (b) two, and (c) six strands of HexCoil-Ala peptide simulated on a (6,5)-SWCNT using REMD. (d) For each residue (30 per strand), the fraction of the time spent in a helix is plotted.

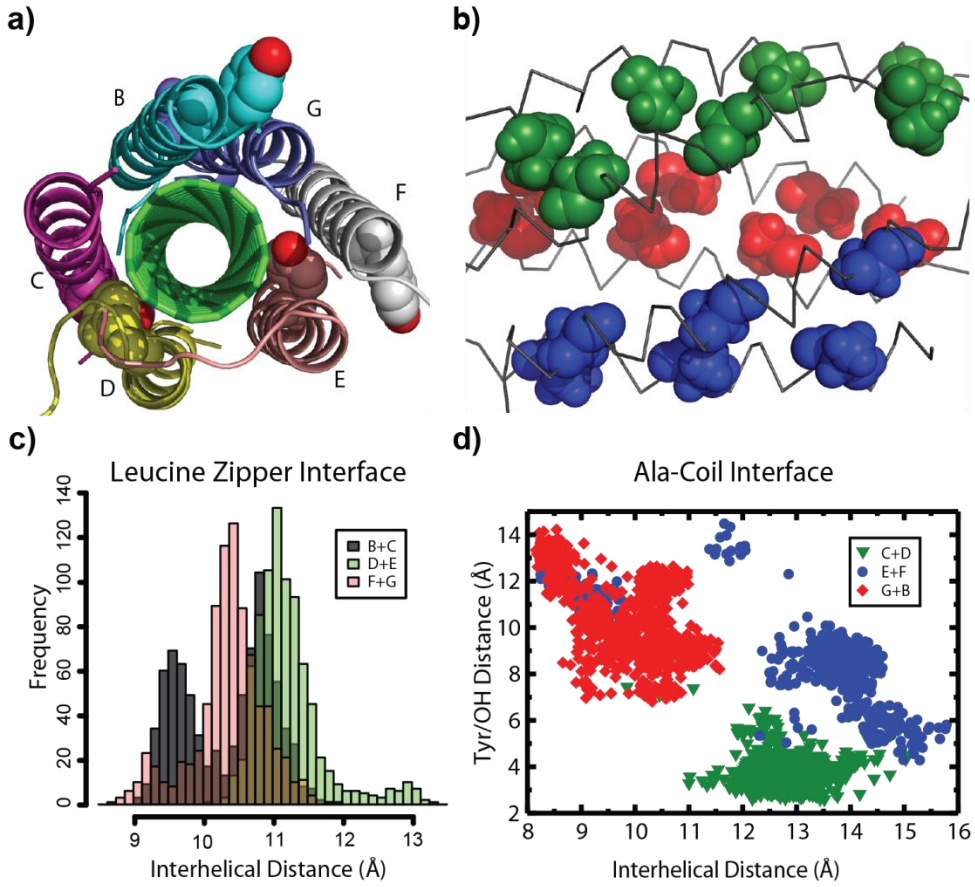


Figure 4. (a) A representative equilibrium structure of the peptide hexamer binding to SWCNT (green), looking down the axis. The chain names of the hexamer are labeled "B" through "G" with tyrosine hydroxyl groups colored red. (b) Side-on view of the hexamer from (a) with alanine residues shown in space-filling representation. Coloring scheme is the same as the legend in (d). (c) Interhelical distance histograms of adjacent peptide dimers. (d) Distance of the hydroxyl group in tyrosine to SWCNT versus interhelical distance in Ala-coil pairs.

3.8 References

1. Sementsov, Y., et al., *Carbon Nanotubes Filled Composite Materials Carbon Nanomaterials in Clean Energy Hydrogen Systems - II*, S.Y. Zaginaichenko, et al., Editors. 2011, Springer Netherlands. p. 183-195.
2. Aikawa, S., et al., *Facile fabrication of all-SWNT field-effect transistors.* Nano Research, 2011. **4**(6): p. 580-588.
3. Nougaret, L., et al., *80 GHz field-effect transistors produced using high purity semiconducting single-walled carbon nanotubes.* Applied Physics Letters, 2009. **94**(24): p. 243505-3.
4. Jin, H., D.A. Heller, and M.S. Strano, *Single-particle tracking of endocytosis and exocytosis of single-walled carbon nanotubes in NIH-3T3 cells.* Nano Letters, 2008. **8**(6): p. 1577-1585.
5. Liu, Z., et al., *Multiplexed multicolor raman imaging of live cells with isotopically modified single walled carbon nanotubes.* J. Am. Chem. Soc., 2008. **130**(41): p. 13540-13541.
6. Welsher, K., et al., *Selective probing and imaging of cells with single walled carbon nanotubes as near-infrared fluorescenct molecules.* Nano Letters, 2008. **8**(2): p. 586-590.
7. Zavaleta, C., et al., *Noninvasive Raman spectroscopy in living mice for evaluation of tumor targeting with carbon nanotubes.* Nano Letters, 2008. **8**(9): p. 2800-2805.
8. Zerda, A., et al., *Carbon nanotubes as photoacoustic molecular imaging agents in living mice.* Nature Nanotechnology, 2008. **3**: p. 557-562.
9. Kostarelos, K., *The long and short of carbon nanotube toxicity.* Nature Biotechnology, 2008. **26**: p. 774-776.
10. Poland, C.A., et al., *Carbon nanotubes introduced into the abdominal cavity of mice show asbestos-like pathogenicity in a pilot.* Nature Nanotechnology, 2008. **3**: p. 423-428.
11. Liu, Z., et al., *In vivo biodistribution and highly efficient tumour targeting of carbon nanotubes in mice.* Nature Nanotechnology, 2007. **2**: p. 47-52.
12. Liu, Z., et al., *Circulation and long-term fate of functionalized, biocompatible single-walled carbon nanotubes in mice probed by Raman spectroscopy.* PNAS, 2008. **105**(5): p. 1410-1415.
13. Liu, Z., et al., *siRNA delivery into human t cells and primary cells with carbon nanotube transporters.* Angew. Chem., 2007. **119**: p. 2069-2073.
14. Bahr, J.L., et al., *Dissolution of small diameter single-wall carbon nanotubes in organic solvents?* Chem. Commun., 2001: p. 193-194.
15. Furtado, C.A., et al., *Debundling and dissolution of single-walled carbon nanotubes in amide solvents.* J. Am. Chem. Soc., 2004. **126**: p. 6095-6105.
16. Huang, X., S.R.L. Mclean, and M. Zheng, *High-Resolution Length Sorting and Purification of DNA-Wrapped Carbon Nanotubes by Size-Exclusion Chromatography.* Analytical Chemistry, 2005. **77**(19): p. 6225-6228.
17. Zheng, M., et al., *Structure-Based Carbon Nanotube Sorting by Sequence-Dependent DNA Assembly.* Science, 2003. **302**(5650): p. 1545-1548.

18. Arnold, M.S., S.I. Stupp, and M.C. Hersam, *Enrichment of Single-Walled Carbon Nanotubes by Diameter in Density Gradients.* Nano Letters, 2005. **5**(4): p. 713-718.
19. Arnold, M.S., et al., *Sorting carbon nanotubes by electronic structure using density differentiation.* Nat Nano, 2006. **1**(1): p. 60-65.
20. Bachilo, S.M., et al., *Structure-assigned optical spectra of single-walled carbon nanotubes.* Science, 2002. **298**: p. 2361-2366.
21. Dieckmann, G.R., et al., *Controlled assembly of carbon nanotubes by designed amphiphilic peptide helices.* J. Am. Chem. Soc., 2003. **125**: p. 1770-1777.
22. Zheng, M., et al., *DNA-assisted dispersion and separation of carbon nanotubes.* Nature Materials, 2003. **2**: p. 338-343.
23. Tu, X., et al., *DNA sequence motifs for structure-specific recognition and separation of carbon nanotubes.* Nature, 2009. **460**(7252): p. 250-253.
24. Zorbas, V., et al., *Preparation and Characterization of Individual Peptide-Wrapped Single-Walled Carbon Nanotubes.* Journal of the American Chemical Society, 2004. **126**(23): p. 7222-7227.
25. Hashida, Y., et al., *Development of a novel composite material with carbon nanotubes assisted by self-assembled peptides designed in conjunction with β-sheet formation.* Journal of Pharmaceutical Sciences: p. n/a-n/a.
26. Khripin, C.Y., et al., *Measurement of Electrostatic Properties of DNA-Carbon Nanotube Hybrids by Capillary Electrophoresis.* Journal of Physical Chemistry C, 2009. **113**(31): p. 13616-13621.
27. Grigoryan, G., et al., *Computational Design of Virus-Like Protein Assemblies on Carbon Nanotube Surfaces.* Science, 2011. **332**(6033): p. 1071-1076.
28. Zuo, G., et al., *Protein Conformational Changes Upon Binding with Carbon Nanotubes.* Curr. Phys. Chem., 2012. **2**(1): p. 12-22.
29. Roxbury, D., et al., *Recognition ability of DNA for carbon nanotubes correlates with their binding affinity.* Langmuir, 2011. **27**(13): p. 8282-8293.
30. Jorgensen, W.L., et al., J. Chem. Phys., 1983. **79**: p. 926.
31. York, D.M., T.A. Darden, and L.G. Pedersen, *The effect of long-range electrostatic interactions in simulations of macromolecular crystals: a comparison of the Ewald and truncated list methods.* J. Chem. Phys., 1993. **99**(10).
32. Humphrey, W., A. Dalke, and K. Schulten, *VMD - visual molecular dynamics.* J. Molec. Graphics, 1996. **14**(1): p. 33-38.
33. Sugita, Y., A. Kitao, and Y. Okamoto, *Multidimensional replica-exchange method for free-energy calculations.* J. Chem. Phys., 2000. **113**: p. 6042-6051.
34. Sugita, Y. and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding.* Chem. Phys. Lett., 1999. **314**(1-2): p. 141-151.
35. Johnson, R.R., et al., *Free Energy Landscape of a DNA-Carbon Nanotube Hybrid Using Replica Exchange Molecular Dynamics.* Nano Letters, 2009. **9**(2): p. 537-541.
36. Roxbury, D., A. Jagota, and J. Mittal, *Sequence-Specific Self-Stitching Motif of Short Single-Stranded DNA on a Single-Walled Carbon Nanotube.* Journal of the American Chemical Society, 2011. **133**(34): p. 13545-13550.

37. Roxbury, D., J. Mittal, and A. Jagota, *Molecular-Basis of Single-Walled Carbon Nanotube Recognition by Single-Stranded DNA.* Nano Letters, 2012. **12**(3): p. 1464-1469.
38. Berendsen, H.J.C., D. van der Spoel, and R. van Drunen, *GROMACS: a message-passing parallel molecular dynamics implementation.* Computer Physics Communications, 1995. **91**(1-3): p. 43-56.
39. Lindahl, E., B. Hess, and D. van der Spoel, *GROMACS 3.0: a package for molecular simulation and trajectory analysis.* J. Mol. Model, 2001. **7**(8): p. 306-317.
40. van der Spoel, D., et al., *GROMACS: fast, flexible, and free.* J. Comput. Chem., 2005. **26**(16): p. 1701-1718.
41. Best, R.B. and G. Hummer, *Optimized Molecular Dynamics Force Fields Applied to the Helix−Coil Transition of Polypeptides.* The Journal of Physical Chemistry B, 2009. **113**(26): p. 9004-9015.
42. Gernert, K.M., et al., *The Alacoil - a Very Tight, Antiparallel Coiled-Coil of Helices.* Protein Science, 1995. **4**(11): p. 2252-2260.
43. Walters, R.F.S. and W.F. DeGrado, *Helix-packing motifs in membrane proteins.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(37): p. 13658-13663.
44. Grigoryan, G. and W.F. DeGrado, *Probing Designability via a Generalized Model of Helical Bundle Geometry.* Journal of Molecular Biology, 2011. **405**(4): p. 1079-1100.

3.9 Supplemental Information

3.9.S1 Circular Dichroism of HexCoil-Ala-SWCNT Samples with Added Surfactant

In this study, we make use of several small surfactant molecules in an attempt to displace HexCoil-Ala peptide from the surface of a preexisting SWCNT suspension. A valid concern would be if the surfactant and peptide interact in a certain manner, causing the reaction to halt. One can imagine that the surfactant, with amphiphilic properties, could bind to the likewise amphiphilic peptide on the SWCNT. Further, if this were to happen, would the secondary alpha-helical structure of the peptide be disrupted? To examine these scenarios, we performed far-UV circular dichroism (CD) spectroscopy on samples of HexCoil-Ala and HexCoil-Ala-SWCNT. To these samples, we have added a small surfactant, sodium dodecyl sulfate (SDS), chosen for its low absorbance in the UV region. This surfactant is similar to SDBS, with the exception of a missing benzyl group (which causes high UV absorbance, and could not be used for CD). Shown in Figure S1, HexCoil-Ala has two negative CD peaks at 208 and 225 nm, indicating the presence of alpha-helical structure. These peaks were not significantly affected by whether or not HexCoil-Ala was wrapping a SWCNT. In addition, SDS also had negligible influence on the two dominant peaks. This suggests that the surfactant is not disrupting the secondary structure of the peptide.

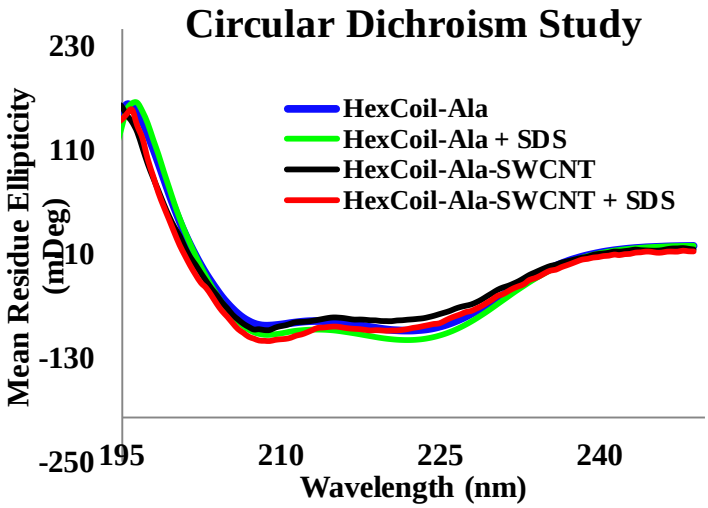


Figure S1. Spectral data from a circular dichroism experiment on HexCoil-Ala and HexCoil-Ala-SWCNT samples illustrating the negligible effect of surfactant, SDS.

3.9.S2 Two-Dimensional Fluorescence Maps

Scanning excitation/emission fluorescence maps can be a very useful tool in determining the quality of a dispersed SWCNT sample. Here we present fluorescence maps from the 'ending' samples of HexCoil-Ala-SWCNT and $(TAT)_4T$-SWCNT after 10 minutes of incubation with SDBS at 60°C (Figure S2). In conjunction with the absorbance scans in Figure 2 of the main text, we can make two conclusions from these fluorescence maps. We can first clearly say that SWCNT emission signals from $(TAT)_4T$ are blue-shifted relative to HexCoil-Ala with.(6,5) emission signal peaks at 980 and 992 nm, respectively. Fluorescence emission in these maps correlates directly with the peak position observed in Figure 2. Therefore, the data confirm that in large measure HexCoil-Ala remains on the SWCNT after an attempted SDBS exchange. The second conclusion that we can draw from the data is that HexCoil-Ala preferentially disperses (6,5)-SWCNT relative to $(TAT)_4T$. Note the intensities of the (8,3) and (7,5)-SWCNTs in both of the maps when normalizing the data by the peak of the (6,5)-SWCNT. The (7,5) peak intensities are 0.46 and 0.83 for HexCoil-Ala and $(TAT)_4T$ dispersions of SWCNT, respectively, showing that HexCoil-Ala preferentially disperses the (6,5)-SWCNT compared to $(TAT)_4T$.

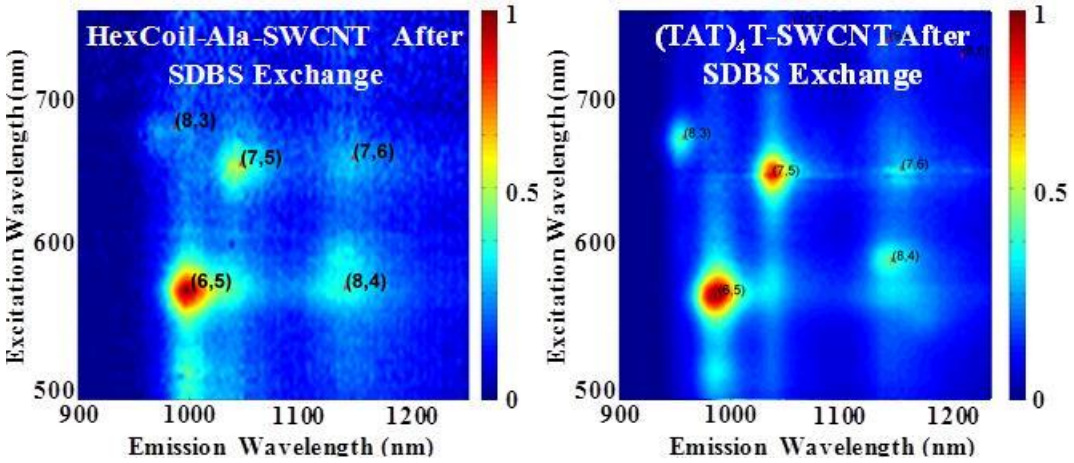


Figure S2. Two-dimensional fluorescence maps for samples of HexCoil-Ala-SWCNT and $(TAT)_4T$-SWCNT after SDBS exchange. Blue-shifted emission wavelengths are seen in the $(TAT)_4T$ sample, suggesting the removal of DNA from the surface of the SWCNTs. Additionally, HexCoil-Ala intensities of SWCNTs other than (6,5) are much lower than in the $(TAT)_4T$ dispersion.

3.9.S3 Convergence in Simulated Structure Analysis

When examining simulations of complex systems, it is useful to estimate the degree of convergence along appropriate reaction coordinates. In the case of simulations of HexCoil-Ala on the (6,5)-SWCNT, we have examined the progressive change in helical fraction in one and two strand configurations. We also observe a sustained amount of helicity in the six-strand configuration. Average helicity versus trajectory for all configurations is shown in Figure S3. It can be seen that the one strand configuration loses a large majority of its helical structure over the course of 60 ns of REMD. Since the two strand configuration is started from two replicas at the end of the one strand, not much change is seen in the average helicity over the same time frame. Additionally, except for an initial drop in the average helicity for each strand, the six-strand configuration remains relatively stable over the course of the simulated trajectory. This further suggests that inter-strand interaction is needed to form stable HexCoil-Ala-SWCNT structures.

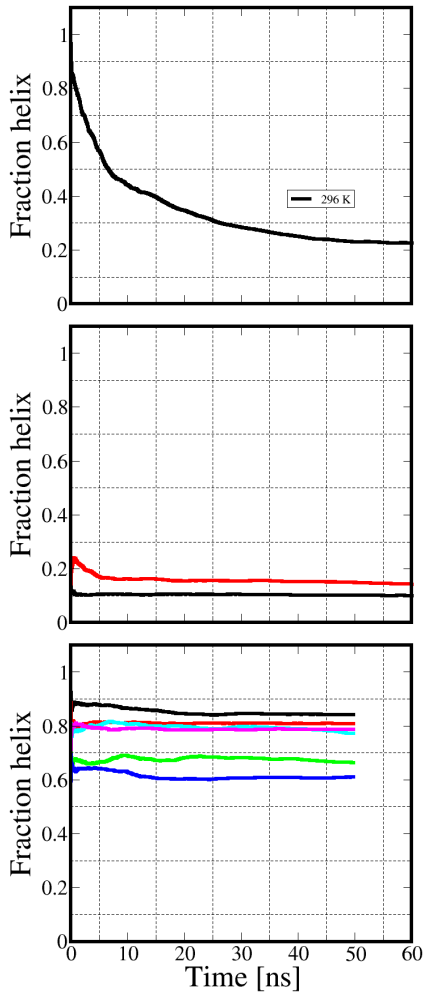


Figure S3. Average helicity vs. time for one, two, and six strand configurations of HexCoil-Ala simulated on a (6,5)-SWCNT. Convergence is determined by a steady value of helicity, generally after an asymptotic decay.

3.9.S4 Characteristic Analysis of the Simulations

When the peptide hexamer binds to SWCNT, its configuration remains stable in simulations. It indicates a sizable affinity between the hexamer and SWCNT. The mean Cα RMSD of the structures in Figure S4a is 1.79 Å. The configuration of the hexamer is different from the designed model by its irregularity. Chain F protrudes from the assembly and has much less contact with SWNCT than the other chains. The local structure of Chain E, F and G resembles that of the tetramer crystal structure. However, the substantial binding between Chains E and G and SWCNT makes a poor alignment in Figure S4b. The mean Cα RMSD on the middle ten residues is 3.20 Å. The stable binding between the hexamer and SWCNT can be easily visualized by compact packing of the peptides on the hydrophobic surface of SWCNT.

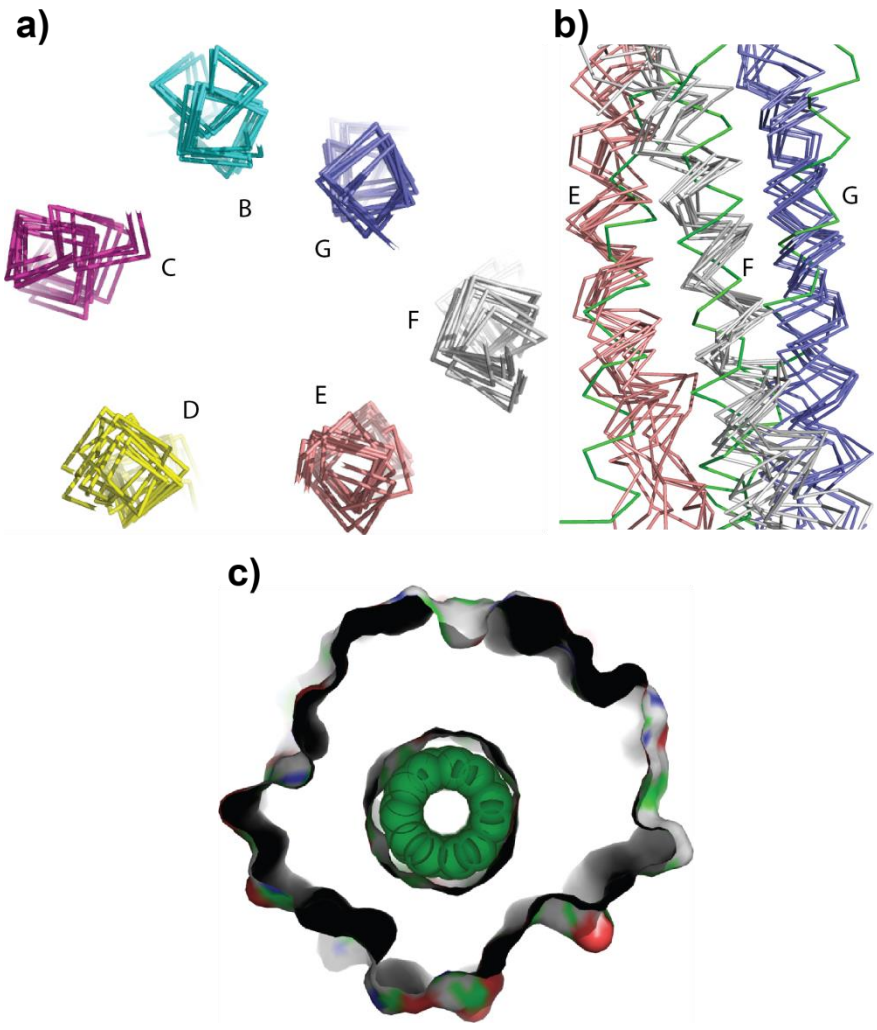


Figure S4. (a) The overlap of the middle ten residues on the configurations from 5 ns to 60 ns with an interval of 5 ns. The chain names are labeled. (b) Chains E, F and G from the structures in (a) are aligned with three consecutive chains from HexCoil-Ala tetramer crystal structure (green) (PDB ID: 3S0R). (c) The surface representation of the peptide hexamer with SWCNT in spheres.

3.9.S5 The Leu-Zipper and Ala-Coil Interfaces

The crystal structure and the designed model have two types of interface, the leucine zipper and Ala-Coil. They have a sequence motif in heptad repeats: L/AXXXXXX, where X stands for any residue. Leucine and alanine are supposed to make close interchain contacts in their specified interfaces as shown in Figure S5.

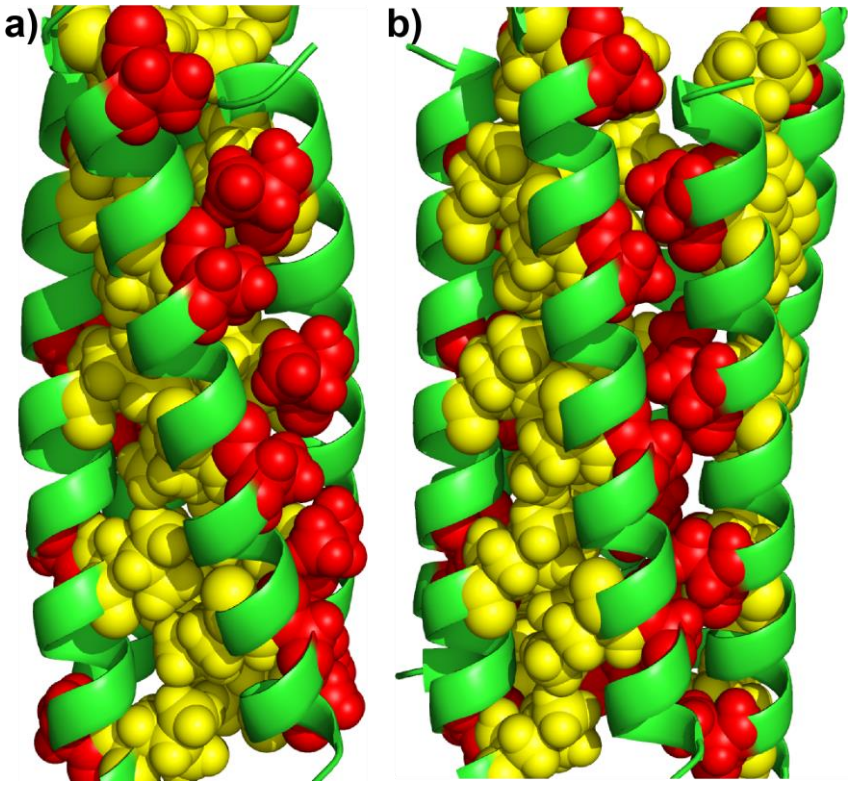


Figure S5. The packing of leucine (yellow) and alanine (red) in the tetramer crystal structure (PDB ID: 3S0R) (a) and the designed hexamer model (b).

Chapter 4

Crystal structure of an amphiphilic foldamer reveals a 48-mer assembly comprising a hollow truncated octahedron

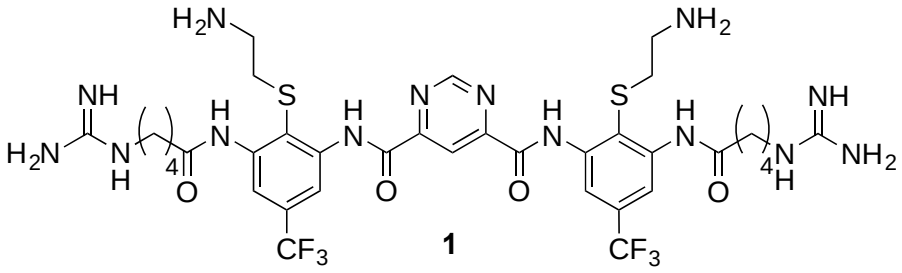
4.1 Overview

The *de novo* design of foldamers provides an approach to test the mechanisms by which biological macromolecules fold into complex three-dimensional structures, and ultimately to design novel protein-like architectures with properties unprecedented in nature. We describe a large cage-like structure formed from an amphiphilic arylamide foldamer crystallized from aqueous solution. Forty eight copies of the foldamer assemble into a 5 nm cage-like structure, an omnitruncated octahedron filled with well-ordered ice-like water molecules. The assembly is stabilized by a mix of arylamide stacking interaction, hydrogen bonding and hydrophobic forces. The omnitruncated octahedra tessellate to form a cubic crystal. These findings provide an important step towards the design of nanostructured particles resembling spherical viruses.

4.2 Introduction

Nature uses a limited set of amino acids to build proteins, which take on a myriad of shapes with defined secondary, tertiary, and quaternary structures. In recent years, chemists have shown that this ability to fold into complex structures is not unique to natural biopolymers, and they have begun building "foldamers" comprised of defined sequences of non-natural building blocks that assemble into increasingly complex secondary structures, and even protein-like folds[1-3]. Moreover, foldamers have been

designed that are responsive to ligand-binding, temperature, or that bind to native biologically important proteins, membranes, and oligosaccharides [4-11]. In this paper, we describe the structural assembly of an amphiphilic arylamide foldamer, **1**, which assembles into a 48-mer cage-like structure.



Compound **1** was originally designed as a mimic of antimicrobial peptides [12]. It is a triarylamide comprising two 1,3-diaminobenzene units linked by a 4,6-dicarboxy-substituted pyrimidine and two terminal guanidine-containing amides. Pendant thioether substituents within the diaminobenzene units help to rigidify the structure and also provide points of attachment for positively charged aminoethyl sidechains. Together, these groups create a facially amphiphilic, positively charged structure, previously shown to be essential for their high antibacterial activity *in vitro* and in animal models [12]. The crystallographic structure of **1** confirms its amphiphilic structure, which is required for binding to bilayers. More importantly, the foldamer, which was crystallized from aqueous solution in the absence of membranes, associated to form in honeycomb geometry with each cubicle as a truncated octahedron. The assembly can be understood in terms of physicochemical principles, including the hydrophobic effect, aromatic stacking, hydrogen bonding, and ion-binding, and should advance the nano-scale engineering of complex molecular assemblies.

4.3 Results

The triarylamide was crystallized from aqueous solution by the hanging drop method in the presence of 0.05 M cadmium acetate and sodium sulfate (1.0 M). Two monomers with closely related structures form the asymmetric unit (Figure 1). Each monomer shows the expected amphiphilic structure, but the arylamide backbone is stabilized differently in the presence than in the absence of Cd^{2+} [13]. In the crystal, Cd^{2+} ions displace the arylamide protons on the amide units connecting the phenyl and pyrimidyl rings. Each Cd^{2+} interacts with the pyrimidyl nitrogen and thioether – replacing the hydrogen-bonded interactions used in the original design with metal-ligand interactions. The neighboring aminoethyl group serves as a fourth ligand, and acetates or (in one case) a water molecule complete the ligand environment. However there are no ligand metal ion interactions between sites within a single molecule or between molecules in the crystal lattice. Thus, the metal ions appear to promote crystallization, not by bridging between sites as in assembly systems formed between polycoordinate metal ions and polydentate ligands [14], but rather by subtly changing the physical and geometric properties of the molecule.

The structure of the triarylamide monomer in the crystallographic lattice displays the amphiphilic structure anticipated in the design. The trifluoromethyl groups and the nonpolar portions of the aryl backbone segregate from the strongly polar amine and guanidine sidechains. The overall arrangement is consistent with that determined by solid-state NMR of the triarylamide bound to phospholipid bilayers in the absence of Cd^{2+} [15]. It is unlikely that the metal ion plays a significant role in the biological activity

of the molecule, because other variants of this triarylamide that lack the thioether and/or the ethylamine ligating groups have high antimicrobial activity [7, 11].

The arylamide crystallizes in a space group of high symmetry, $P\ \bar{4}\ 3\ n$, which to the best of our knowledge has not yet been seen for an organic molecule. The unit cell (Figure 2-3) contains 24 copies of the asymmetric dimer, arranged in the shape of an omnitruncated octahedron, a special case of an Archimedean solid, the truncated octahedron, which has eight hexagonal and six square faces. Each arylamide dimer associates with 131 water molecules, which are primarily located within the cores of the truncated octahedra, for a total of 3144 water molecules per unit cell.

The 48 triarylamides lie with their backbones roughly parallel to the edges of the eight hexagons (Figure 2). Each arylamide engages in two distinct types of interactions, which together uniquely define and stabilize the overall assembly. Arylamides that lie along the edges of neighboring hexagons interact in what we term a "trifluoromethyl zipper" interaction, in which the water-repelling trifluoromethyl groups intimately interdigitate along a non-exact two-fold axis of symmetry. While engaging in trifluoromethyl zipper interactions, the arylamides also engage in a second type of interaction between arylamides that lie within a single hexameric ring. We designate this interaction the arylamide elbow motif; the terminal phenyl rings of the arylamides in this motif stack in a face-to-face interaction with the centers of the rings offset as often seen in aromatic stacking [16]. The dimer is further stabilized by tight hydrogen bonding between the terminal amide groups, and their trifluoromethyl groups also cluster together. The

hydrogen-bonded interaction between meta-substituted amides leads to a 120° angle between the two arylamides, which is repeated to create the hexameric rings.

The simultaneous interactions of the triarylamides in both the trifuoromethyl zipper and the arylamide elbow motif create the 48-mer assembly seen in the crystal (Figure 2). The arylamide elbows create the hexamers, while the trifluoromethyl zippers couple adjacent hexamers to form the overall three-dimensional structure. Interestingly, this arrangement also leads to extensive clustering of the trifluoromethyl groups along the vertices, creating fluorocarbon cores (Figure 3A).

Electrostatic and hydrogen-bonded interactions between sulfate ions and the arginine-like guanidine sidechains feature prominently in each four-sided face of the truncated octahedral (Figure 3). A total of four sulfates are seen in each face, two coming from each unit cell of the crystal lattice (Figure 3A). Each sulfate receives a total of six hydrogen bonds from three guanidine-containing sidechains that engage the anion in a bidentate interaction. The hexagonal faces also show a rich array of molecular interactions; the methyl groups of the acetates (counterions of the Cd^{2+} ions) project towards the center of the hexagon (Figure 3B). Each acetate ion is surrounded by a clathrate of water molecules, which join near the center to create a nearly ideal ice-like hexagon of water molecules.

The individual truncated octahedra found in the unit cell tessellate in three dimensions to form a cube. Intersecting square and hexagonal channels run through the length of the crystal (Figure 4). Arylamides that do not form a trifluoromethyl zipper within an individual unit cell, form an equivalent zipper motif with other monomers from adjacent

unit cells. Also, the packing of the truncated octahedra contribute to the cluster of sulfate/guanidine interactions along the square channels, as well as the clustering of trifluoromethyl groups at the vertices and along the hexagonal channels. Thus, the same interactions that stabilize individual truncated octahedra also contribute to their packing into a crystal lattice.

To determine how the compound behaves in solution, we also performed NMR experiments in a buffer similar to the one in which the crystal was obtained. Titration of the compound with Cd^{2+} ions in Figure S1 shows that upon addition of metal ion the peaks were broadened and in some cases appear at new positions in the spectrum. At sub-stoichiometric Cd^{2+} concentrations, multiple peaks are observed, indicating that different forms of **1** with 0, 1, and 2 equivalents of Cd^{2+} bound were in slow to intermediate exchange. After a stoichiometric amount (two equivalents per foldamer) of Cd^{2+} is reached the peaks begin sharpening as the distribution becomes more homogeneous. Little change is observed after 5.0 equivalents (40 mM) were added. These data show that the foldamer binds Cd^{2+} stoichiometrically at this concentration, and that it existed in the Cd^{2+}-bound state under conditions of crystallization.

To examine the self-association of the foldamer we examined the concentration dependence of its proton and ^{19}F NMR spectrum, while holding the Cd^{2+} constant at saturating concentrations. As the concentration was increased to 2 mM the proton NMR spectrum showed small changes, while at 8.2 mM the peaks shifted further and two new peaks appeared in the spectrum (Figure 5A). The new peaks are assigned to amides and guanidine protons from **1**, as was confirmed by hydrogen-deuterium exchange (Figure

5B). The fact that, at the highest concentration, the amides were observable at pH 7.5 in H_2O shows that their exchange with bulk solvent is slowed by the formation of an oligomer. The concentration dependent ^{19}F NMR spectra also showed biphasic changes (Figure 5C). In this case, larger changes in chemical shift were observed between 0.1 and 2 mM. A second process occurs at higher concentrations (between 2 and 10 mM), causing the peaks to shift in the reverse direction and to broaden. While a quantitative analysis is complicated by the many equilibria, these data clearly show that **1** associates in aqueous solution, and that the mean association state increases as the concentration increases.

4.4 Discussion

Proteins are built up from secondary structures with pronounced facially amphiphilic character, which have also served as building blocks in the design of the first helical bundles composed of peptides synthesized from both α- and β-amino acids. Therefore, it was of considerable interest to determine the structures formed by the present triarylamides. The discovery of a large cage-like structure is interesting, given the complex protein-like interactions that stabilize the assembly.

Previous designs of cage-like structures in solution and crystals have generally focused on assemblies built from proteins[17-19] or small molecules[14]. Organic crystals are often assembled through strong directional intermolecular forces in organic solvents [20]. One of the most noteworthy fields in crystal engineering is metal-organic frameworks, assembly systems with predetermined directionalities between polycoordinate metal ions

and polydentate ligands [14]. Additionally, certain organic frameworks are built by strong hydrogen bonding [21].

The assembly described here uses a diversity of interactions similar to those employed by natural proteins. Nevertheless, it is interesting to consider its construction in terms of "supramolecular synthons", as in other examples of crystal engineering. The arylamide elbow motif in triarylamide can be considered a supramolecular synthon, which engages in both hydrogen bonding and aromatic stacking (Figure 1A, in pink shadow). The two N-H···O=C hydrogen bonds in the arylamide elbow motif have a distance of 2.8 Å and an angle of 157º and 166º, which show that they are strong hydrogen bonds as in other examples of crystal engineering [1]. The distance between the two phenyl rings is 3.7 Å, as observed in other crystal studies [22]. Moreover, the pronounced facially amphiphilic character and high symmetry of **1** gives rise to its assembly and crystallization into an omnitruncated octahedron in aqueous environments. The assembly can be conceptually analyzed according to Aufbau principles [23], although the complexity of its structure and self-association equilibria prior to crystallization render it difficulty to assign a detailed kinetic mechanism of assembly. Hydrogen bonding and aromatic stacking provide geometrically specific interactions that stabilize the aromatic elbow (Figure 6A), and repetition of this interaction pattern leads to assembly of the hexameric ring (Figure 6B) that forms the hexagonal face of the truncated octahedron. Two hexagons can further assemble via hydrophobic association of trifluoromethyl groups and hydrogen bonding between the guanidine and carbonyl of neighboring arginine-like groups stabilize (Figure 6C). The burial of trifluoromethyl core is further consolidated by assembly of two more hexagons; the resulting saddle-shaped tetramer of hexamers forms a complete vertex for

the truncated octahedron (Figure 6D). Two such tetramers of hexamers associate and create a “double crown” assembly on the square face (Figure 6E), which serves as a long-range synthon module for packing in the late stages of crystallization [24]. This supramolecular subunit packs in primitive cubic cells and generates a crystalline framework (Figure 6F).

The serendipitous discovery of a framework structure has potential implications for the rational design of nanoporous solids in aqueous environments. The interactions that stabilize this visually arresting Archimedean solid are readily apparent and should encourage future rational designs of related structures.

4.5 Methods

Crystallography The arylamide foldamer **1** was synthesized as described previously [12], and dissolved at 20 mg/ml in water. All the crystals were obtained using the hanging-drop vapor-diffusion method at room temperature, by mixing equal volumes of foldamer solution with reservoir solution containing the crystallization reagent. Crystallization conditions were determined by biased sparse matrix crystallization screen (Hampton Research). 0.05 M CdSO4, 0.1 M HEPES pH 7.5 and 1.0 M NaOAc (Crystal Screen 2, #34) gave diffraction quality crystals, which were flash frozen with the cryoprotectant Parabar 10312 (Hampton Research). The diffraction data were collected on the beamlines 24-ID-E and 24-ID-C at Advanced Photon Source (APS) at Argonne National Laboratory. The best crystal was diffracted to 0.920 Å, the highest resolution limit achievable at the beam-line. Data were processed using the HKL2000 [25] software. The resolution range 55.00 to 0.960 Å was utilized to solve the crystal structure. The

overall completeness of the data is 99.8% in this resolution shell. A total of 29271 reflections were measured, while 15265 reflections were used after merging equivalent reflections. The overall R_{merge} is 0.012, whereas the redundancy is as high as 31.8.

The formula of the foldamer crystal is $2(C_{36}H_{46}N_{14}O_4F_6S_2Cd^{2+}{}_2)\cdot 7(CH_3COO^-)\ Na^+\cdot SO_4^{2-}\cdot 134.42H_2O$, MW=5237.52 a.m.u., cubic, space group $P\ \bar{4}\ 3\ n$ (no. 218), Z = 24, a = 53.063(1) Å , V = 149409(5) $Å^3$, D_x = 1.397 g cm^{-3}. The structure has been solved by direct methods by using the program SIR2008 [26]. The phase set with the best figure of merit (best final FoM = 2.589) allowed to identify most of the core atoms of the arylamide foldamer. The structure was completed by the use of the program CRYSTALS [27]. Fourier analysis revealed the presence of one sulfate ion and one sodium ion. The preliminary structure model showed the presence of empty channels and subsequent Fourier and least-squares cycles highlighted the presence of additional water molecules. A total of 138 water molecules were identified. One water molecule is coordinated to a cadmium ion, five hydrate the sodium ion, and two are statistically placed at two different positions. Two water molecules lie on the 3-fold axis, one on the 2-fold axis, and one on the 4-fold axis. Anti-bumping restraints have been used for water molecules. No contribution of diffuse solvent was used. Trifluoromethyl groups show high thermal motion and two out of four trifluoromethyl groups were split into two different staggered conformations. In addition, two out of four aminoethyl groups show two different staggered conformations. For clarity, only one rotamer for the aminomethyl and trifluoromethyl groups is shown in the figures. Atom occupancy, related to the statistically distributed conformations, and statistically placed water molecules, was also refined. Anisotropic thermal factors were used only for all non-hydrogen atoms of the

arylamide foldamer. A total of 2081 refinable parameters were finally considered. Mogul geometry check was performed, and a total of 188 distance restraints were applied. Thermal vibration and thermal similarity restraints were also used. DIFABS [28] absorption correction was applied to the data during the refinement. Hydrogen atoms were geometrically placed to the carrying atoms and ride during refinement. Water molecule hydrogen atoms were not included in the refinement. Chebychev polynomial [29, 30] weighting scheme was used. Final disagreement index considering 13724 independent reflections with $I \geq 3.0\sigma(I)$ is $R_1 = 0.0876$, while $w_R = 0.1051$ for 14550 reflections with $I \geq 1.0\sigma(I)$, and $S = 1.099$. The absence of residual density in the lattice voids has been also verified by means of difference Fourier maps.

NMR spectroscopy All spectra were recorded at 298 K on a Bruker 900 MHz spectrometer equipped with a cryogenic probe for ^{1}H spectra or a Bruker 300 MHz for ^{19}F spectra. ^{1}H spectra typically were recorded with 256 scans and 31 ppm spectral width and ^{19}F spectra typically were recorded with 256 scans and 100 ppm spectral width. ^{1}H chemical shifts were referenced with respect to the residual water peak at 4.63 ppm and the ^{19}F-chemical shifts were calibrated using the external standard trifluoroacetic acid chemical shifts at -76.6 ppm. All spectra were processed and analyzed using the programs TopSpin 3.0. Prior to Fourier transformation, time domain data were multiplied by sine square bell window functions shifted by 90º and zero-filled once.

4.6 Acknowledgments

Vincenzo Pavone and Yibing Wu are the co-authors of this Chapter. I thank Nathan Joh for technical assistance. This work was supported by NIH grant GM54616, and the MRSEC program of NIH through a grant to the LRSM at the University of Pennsylvania.

4.7 Figures

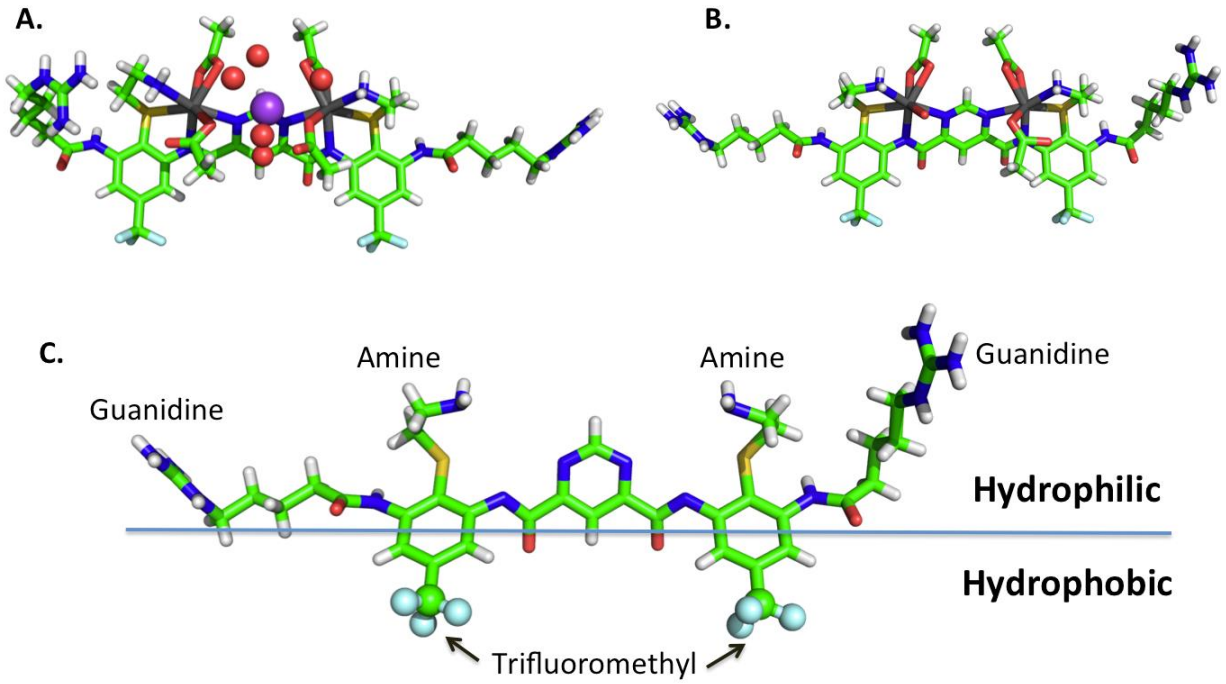


Figure 1. (**A&B**) Structure of two monomers in the asymmetric unit. Monomer 1 (**A**) contains two acetates bound to each Cd(II) ion (grey). A hydrated Na^+ ion (purple sphere) binds between the two Cd(II) sites, maintaining electrical neutrality. Monomer 2 (panel **B**) has two acetates bound to one Cd(II), a single acetate bound to the second Cd(II) and no sodium ion. The metal-binding sites are shown in more detail in figure S1. Panel **C** shows monomer 1 with the cadmium, sodium, and acetate ions removed.

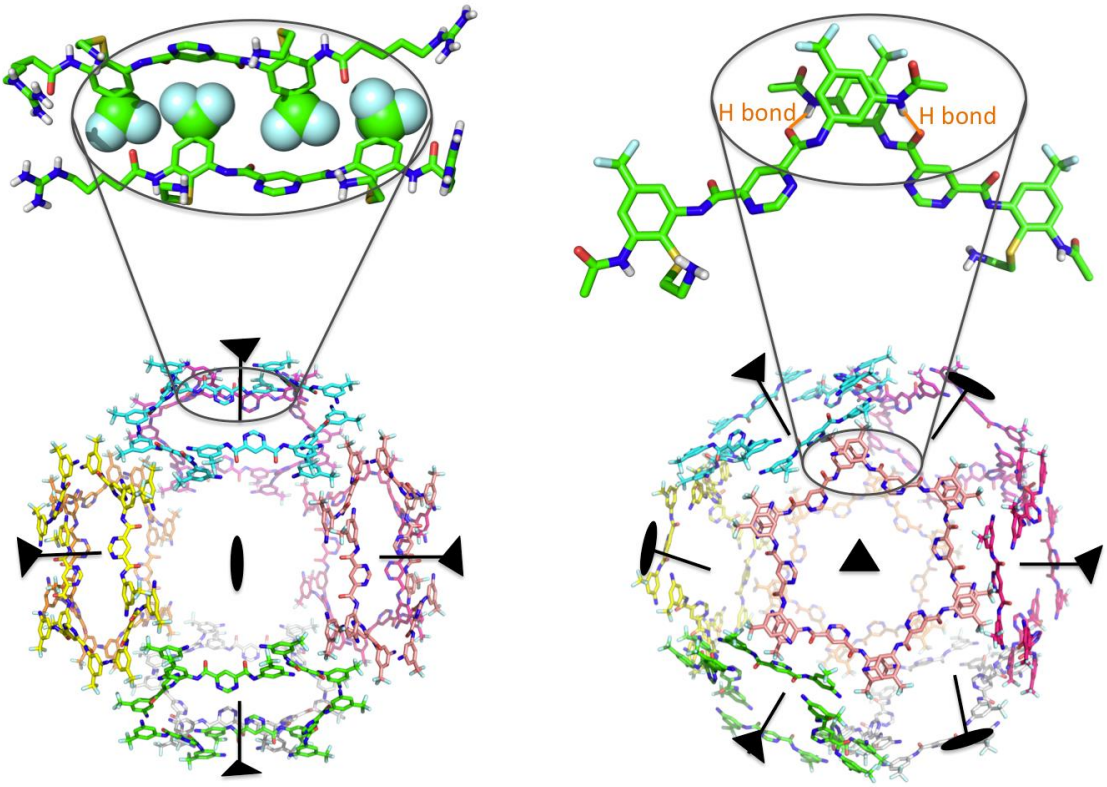


Figure 2. Omnitruncated octahedron formed by 48 copies of the triarylamide. At left, the structure is viewed down the square faces, which have two-fold symmetry. The hexagonal faces have crystallographic 3-fold symmetry as shown. The inset shows the packing between arylamide neighbors interacting at edges between hexagonal faces. At right the structure is viewed down the hexagonal face, which has three-fold crystallographic symmetry. The inset shows that the individual arylamides interact with their aryl groups stacked in an offset manner, the adjacent terminal amides in a tight hydrogen bond, and the trifluoromethyl groups in close proximity.

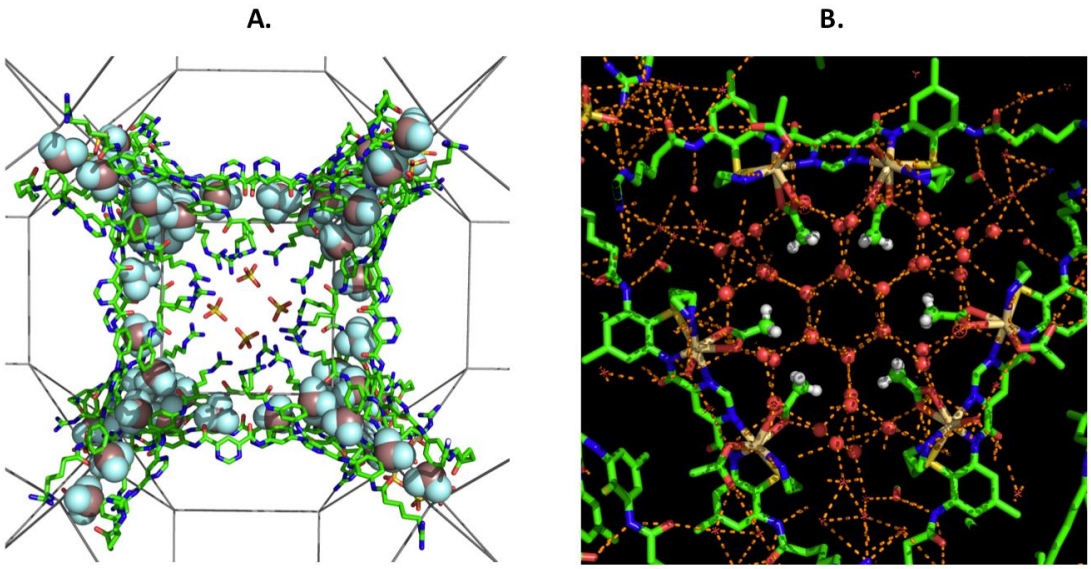


Figure 3. Interactions of small molecules and water with the triarylamide along the square (A) and hexagonal (B) faces. In panel A, sulfate ions are shown interacting with the Arg-like sidechain of the arylamides. Trifluorometyl groups are also shown in space-filling representation. In (B) the acetates are shown with their methyl groups (protons included) surrounded by a clathrate of water molecules.

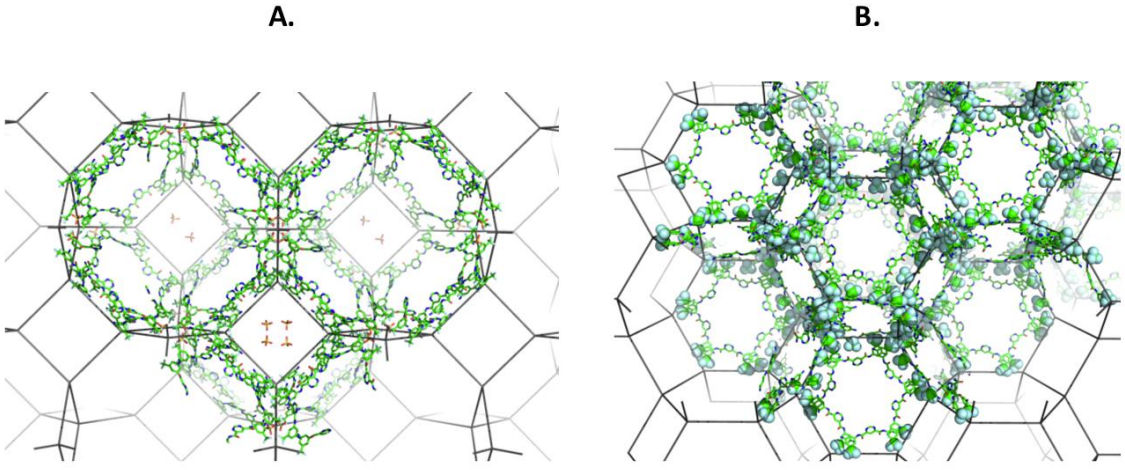


Figure 4. Packing of the unit cells in the crystal structure, viewed down the square and hexagonal channels.

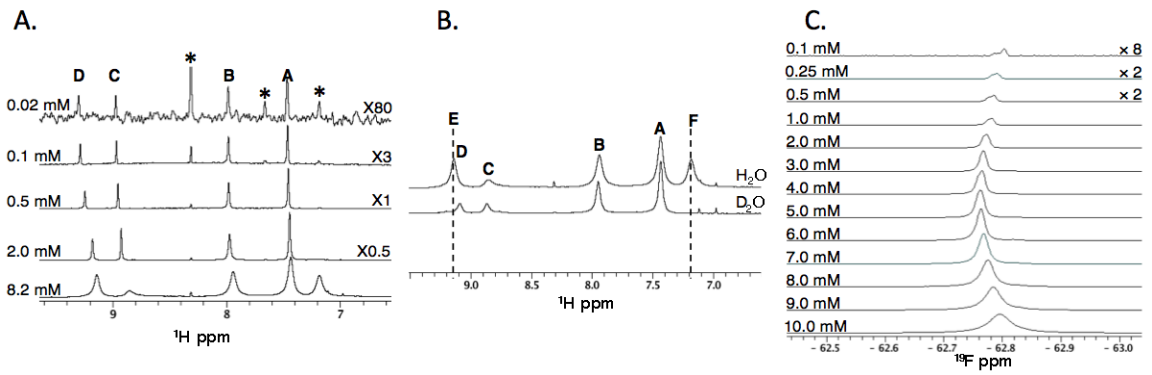


Figure 5. (A) ^{1}H NMR spectra as a function of the concentration of **1** (298 K, 95% H_2O/5% D_2O, 30 mM $CdSO_4$, 100 mM HEPES pH 7.5, 600 mM NaOAc). (B) Comparison of the same sample recorded in H_2O and D_2O shows the disappeared exchangeable protons **E** and **F** in D_2O highlighted by the dash lines. The H_2O sample was 8.2 mM compound in 30 mM cadmium sulfate, 100 mM HEPES pH 7.5, 600 mM sodium acetate and the D_2O sample was obtained by lyophilizing the H_2O sample overnight and then adding D_2O. (C) ^{19}F NMR spectra as a function of concentration of **1** at 95% H_2O/5% D_2O, 100 mM HEPES pH 7.5, 1 M NaOAc, 50 mM $CdSO_4$. Assignments for protons **A**-**F** are given in the Supplemental Figure S1. Peaks labeled with a * are impurities in the buffer.

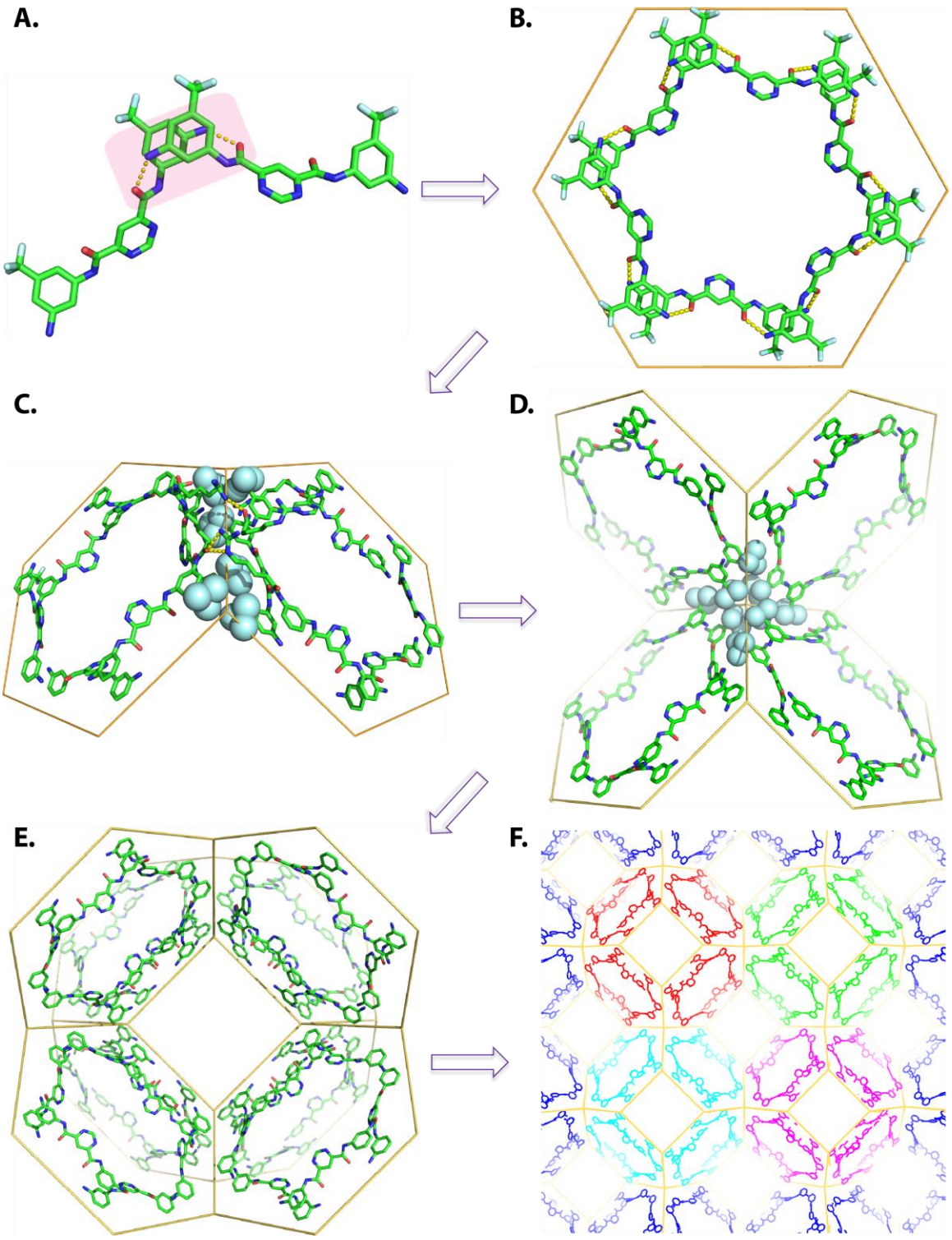


Figure 6. Hierarchic assembly of the foldamer crystal. Hydrogen bonds are shown as yellow dashed lines and fluorine atoms are displayed as spheres. The yellows wires are

lattices defined by the honeycomb symmetry of the crystal. In (A) the supramolecular synthon is denoted in pink shadow.

4.8 Supplemental Figures

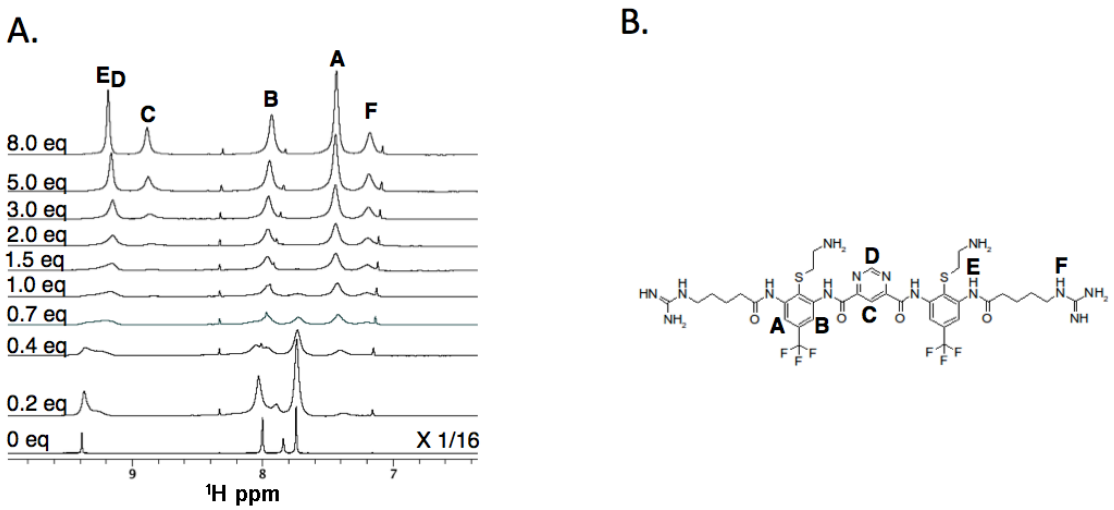


Figure S1. (A) ^{1}H NMR spectra of the foldamer compound (monomer concentration at 8.2 mM) at 298 K, 95% H_2O/5% D_2O, 100 mM HEPES pH 7.5, 600 mM NaOAc, titrated by addition of small aliquots from a concentrated stock solution (800 mM) $CdSO_4$. The metal to compound ratio is labeled on the left of each spectrum. (B) Chemical structure of the compound with protons labeled **A**-**F**.

4.9 References

1. Goodman, J.L., et al., *Tetrameric beta(3)-peptide bundles.* Chembiochem, 2008. **9**(10): p. 1576-8.
2. Petersson, E.J. and A. Schepartz, *Toward beta-amino acid proteins: design, synthesis, and characterization of a fifteen kilodalton beta-peptide tetramer.* J Am Chem Soc, 2008. **130**(3): p. 821-3.
3. Cheng, R.P., S.H. Gellman, and W.F. DeGrado, *beta-Peptides: from structure to function.* Chem Rev, 2001. **101**(10): p. 3219-32.
4. Huc, I., *Aromatic oligoamide foldamers.* European Journal of Organic Chemistry, 2004(1): p. 17-29.
5. Hecht, S. and I. Huc, eds. *Foldamers; Structure, Properties, and Applications*. 2007, Wiley-VCH: Weinheim, Germany.
6. Lee, E.F., et al., *Structural basis of Bcl-xL recognition by a BH3-mimetic alpha/beta-peptide generated by sequence-based design.* Chembiochem, 2011. **12**(13): p. 2025-32.
7. Tew, G.N., et al., *De novo design of antimicrobial polymers, foldamers, and small molecules: from discovery to practical applications.* Acc Chem Res, 2010. **43**(1): p. 30-9.
8. Qi, T., T. Deschrijver, and I. Huc, *Large-scale and chromatography-free synthesis of an octameric quinoline-based aromatic amide helical foldamer.* Nat Protoc, 2013. **8**(4): p. 693-708.
9. Horne, W.S. and S.H. Gellman, *Foldamers with Heterogeneous Backbones.* Acc Chem Res, 2008.
10. Goodman, C.M., et al., *Foldamers as versatile frameworks for the design and evolution of function.* Nat Chem Biol, 2007. **3**(5): p. 252-62.
11. Scott, R.W., W.F. DeGrado, and G.N. Tew, *De novo designed synthetic mimics of antimicrobial peptides.* Curr Opin Biotechnol, 2008. **19**(6): p. 620-7.
12. Choi, S., et al., *De novo design and in vivo activity of conformationally restrained antimicrobial arylamide foldamers.* Proc Natl Acad Sci U S A, 2009. **106**(17): p. 6968-73.
13. Tang, H., et al., *Biomimetic facially amphiphilic antibacterial oligomers with conformationally stiff backbones.* Chem Biol, 2006. **13**(4): p. 427-35.
14. Farrusseng, D., *Metal-organic frameworks : applications from catalysis to gas storage*. 2011, Weinheim: Wiley-VCH. xxii, 392 p.
15. Su, Y., W.F. DeGrado, and M. Hong, *Orientation, dynamics, and lipid interaction of an antimicrobial arylamide investigated by 19F and 31P solid-state NMR spectroscopy.* J Am Chem Soc, 2010. **132**(26): p. 9197-205.
16. Meyer, E.A., R.K. Castellano, and F. Diederich, *Interactions with aromatic rings in chemical and biological recognition.* Angew Chem Int Ed Engl, 2003. **42**(11): p. 1210-50.
17. King, N.P., et al., *Computational design of self-assembling protein nanomaterials with atomic level accuracy.* Science, 2012. **336**(6085): p. 1171-4.
18. Lanci, C.J., et al., *Computational design of a protein crystal.* Proc Natl Acad Sci U S A, 2012. **109**(19): p. 7304-9.

19. Lai, Y.T., et al., *Structure and flexibility of nanoscale protein cages designed by symmetric self-assembly.* J Am Chem Soc, 2013. **135**(20): p. 7738-43.
20. Desiraju, G.R., *Crystal engineering: a holistic view.* Angew Chem Int Ed Engl, 2007. **46**(44): p. 8342-56.
21. Desiraju, G.R., *Crystal design : structure and function*. Perspectives in supramolecular chemistry. 2003, Chichester, West Sussex, England ; Hoboken, NJ: Wiley. xi, 408 p.
22. Chartrand, D., I. Theobald, and G.S. Hanan, *Bis[4′-(3,5-dibromophenyl)-2,2′:6′,2″-terpyridine]ruthenium(II) bis(hexafluorophosphate) acetonitrile disolvate.* Acta Crystallographica Section E Structure Reports Online, 2007. **63**(6): p. m1561-m1561.
23. Desiraju, G.R., *Crystal Engineering: From Molecule to Crystal.* J Am Chem Soc, 2013.
24. Ganguly, P. and G.R. Desiraju, *Long-range synthon Aufbau modules (LSAM) in crystal structures: systematic changes in C6H6−nFn (0 ≤n≤ 6) fluorobenzenes.* CrystEngComm, 2010. **12**(3): p. 817.
25. Otwinowski, Z. and W. Minor, *Processing of X-ray diffraction data collected in oscillation mode.* Macromolecular Crystallography, Pt A, 1997. **276**: p. 307-326.
26. Burla, M.C., et al., *IL MILIONE: a suite of computer programs for crystal structure solution of proteins.* Journal of Applied Crystallography, 2007. **40**: p. 609-613.
27. Betteridge, P.W., et al., *CRYSTALS version 12: software for guided crystal structure analysis.* Journal of Applied Crystallography, 2003. **36**: p. 1487-1487.
28. Walker, N. and D. Stuart, *An Empirical-Method for Correcting Diffractometer Data for Absorption Effects.* Acta Crystallographica Section A, 1983. **39**(Jan): p. 158-166.
29. Watkin, D., *The Control of Difficult Refinements.* Acta Crystallographica Section A, 1994. **50**: p. 411-437.
30. Prince, E., *Mathematical techniques in crystallography and materials science*. 1982, New York: Springer-Verlag. viii, 192 p.